APPARATUS AND METHODS FOR AN INFORMATION RETRIEVAL SYSTEM THA EMPLOYS NATURAL LANGUAGE PROCESSING OF SEARCH RESULTS TO IMPROVE OVERALL PRECISION

Patent number: JP2001511564T

Publication date: 2001-08-14

Inventor:

Applicant:

Classification:

- international: ***G06F17/30;*** **G06F17/30;** (IPC1-7): G06F17/30

- european: G06F17/30T2P4N; G06F17/30T

Application number: JP20000504525T 19980513

Priority number(s): US19970898652 19970722; WO1998US09711 19980513

Also published as:

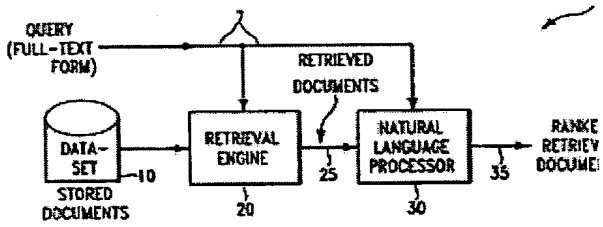
WO9905618 (A
EP0996899 (A1
US6901399 (B
US5933822 (A
EP0996899 (A

more >>

Abstract not available for JP2001511564T

Abstract of correspondent: **WO9905618**

Apparatus and accompanying methods for an information retrieval system that utilizes natural language processing to process results retrieved by, for example, an information retrieval engine such as a conventional statistical-based search engine, in order to improve overall precision. Specifically, such a search ultimately yields a set of retrieved documents. Each such document is then subjected to natural language processing to produce a set of logical forms. Each such logical form encodes, in a word-relation-word manner, semantic relationships, particularly argument and adjunct structure, between words in a phrase. A user-supplied query is analyzed in the same manner to yield a set of corresponding logical forms therefor. Documents are ranked as a predefined function of the logical forms from the documents and the query. Specifically, the set of logical forms for the query is then compared against a set of logical forms for each of the retrieved documents in order to ascertain a match between any such logical forms in both sets. Each document that has at least one matching logical forms is heuristically scored, with each different relation for a matching logical forms being assigned a different corresponding predefined weight. The score of each such document is, e.g., a predefined function of the weights of its uniquely matching logical forms. Finally, the retained documents are ranked in order of descending score and then presented to a user in that order.



* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]Are a device for using a document memorized in an information retrieval system for searching from a repository, and said system, Answer a query, search two or more documents relevant to the query memorized, have a search system for specifying an output document group, and said device, An implication and a processor answer a command memorized by memory in a processor and a memory a excecutable command is remembered to be, a query is answered, the 1st logical form for it is produced, and the 1st logical form shows semantic relations between words relevant to a query -- a document in an output document group -- each -- to one [another], Acquire the 2nd corresponding logical form and the 2nd logical form semantic relations between words relevant to a phrase in said one document An example and the 1st logical form of a query, As a function with the 2nd logical form for one each of two or more documents in an output document group defined beforehand, two or more documents in an output document group are ranked, and the order of a rank is specified -- a device which gives two or more entries relevant to an output document group memorized as an output in order of said rank.
[Claim 2]The device according to claim 1 which each entry is one of the correspondences of a document in an output document group, or is a record relevant to one document of said correspondence.
[Claim 3]the inside of the 1st logical form for a query, and an output document group -- each -- the device according to claim 2 whose each with the 2nd logical form for another document is a list of a logical form graph, its subgraph, or logical form triads, respectively.
[Claim 4]a processor answers a command memorized -- a document in an output document group -- said -- each -- for one [another], or it reads the 2nd logical form of correspondence from a storage -- or -- the inside of an output document group -- said -- each -- the device according to claim 3 which generates the 2nd logical form of said correspondence by analyzing

one another document.
[Claim 5]Said function generates a score based on a relation between each of said 1st logical form relevant to a query, and said 2nd logical form relevant to said one document defined beforehand for one of said the documents, The device according to claim 4 which a processor answers a command memorized, ranks an entry memorized according to a score relevant to each document in an output document group, and specifies the order of a rank.
[Claim 6]said 1st logical form relevant to a query, or said 2nd logical form relevant to one of said the documents in an output document group -- respectively -- said query -- or the device according to claim 5 which contains further a paraphrase of words and phrases relevant to one of said the documents.
[Claim 7]The device comprising according to claim 6:
Each of said 1st logical form and said 2nd logical form One or more logical form triads, Said logical form triad in said 1st list and said logical form triad in said 2nd list including a list and a list [the 2nd] of the 1st of correspondences respectively, A form of each stem of a word of two words which were semantically related to a logical form graph of correspondence in one phrase each of said document in a query, respectively.
A relation showing semantic relations between two words specified beforehand.

[Claim 8]The device according to claim 5 which is coincidence with said same coincidence between arbitrary things of said 1st logical form relevant to a query, and said 2nd logical form relevant to arbitrary documents in an output document group.
[Claim 9]The device comprising according to claim 8:
Each of said 1st logical form and said 2nd logical form includes a list and the 2nd list of the 1st, [correspondence of one or more logical form triads] A form of each stem of a word of two words to which said logical form triad in said 1st list and said logical form triad in said 2nd list were semantically related respectively in a logical form graph of correspondence in one phrase each of said document in a query.
A relation showing semantic relations between two words specified beforehand.

[Claim 10]The device according to claim 5 with which a repository contains a data set.
[Claim 11]The device according to claim 5 whose query is a query of the full text.
[Claim 12]The device according to claim 5 with which a search system contains a statistical search engine.
[Claim 13]A client computer for acquiring a query from a user and displaying two or more documents in an output document group in order of said rank, As for said server, a processor answers a command memorized by memory including said processor and said memory, including further a server connected to a client computer via network connection, a query is

acquired from a client computer -- the device according to claim 5 which gives said two or more documents in a set of an output document to a client computer in order of said rank.
[Claim 14]The device according to claim 13 with which said server contains two or more individual servers.
[Claim 15]The device according to claim 13 with which a search system contains a statistical search engine.
[Claim 16]The device according to claim 15 whose network connection is the Internet or an Internet connectivity.
[Claim 17]A search engine answers a query and a record memorized from a repository for one each of said two or more of the documents in a set of an output document is searched, A processor answers a command memorized by memory and information included in a record including information which pinpoints a place where one each of said the documents in an output document group may be found out as for a record, The device according to claim 16 which accesses one each of said the documents from a server of relation for it, downloads it, and is included in an output document group.
[Claim 18]A client computer which has said processor and said memory, A server connected to a client computer via network connection is included further, The device according to claim 5 which said server realizes said search system, answers a query given by a client computer, and gives said output document group to a client computer.
[Claim 19]The device according to claim 18 with which a search system contains a statistical search engine.
[Claim 20]The device according to claim 19 whose network connection is the Internet or an Internet connectivity.
[Claim 21]A search engine answers a query and a record memorized from a repository for one each of said two or more of the documents in a set of an output document is searched, A processor answers a command memorized by memory and information included in a record including information which pinpoints a place where one each of said the documents in an output document group may be found out as for a record, The device according to claim 20 which accesses one each of said the documents from a server of relation for it, downloads it, and is included in an output document group.
[Claim 22]The device according to claim 5 with which a computer realizes said search system according to a command memorized by memory again including further a computer which has said processor and said memory.
[Claim 23]The device according to claim 22 with which a search system contains a statistical search engine.
[Claim 24]Again a score for said one document A node word in the 2nd [for said one document] logical form, Frequency or semantic contents of said node word in said one

document, The device according to claim 5 which is the function as which frequency of a specific logical form triad for frequency of a node word in said one document specified beforehand or semantic contents, and said one document or the length of said one document was determined beforehand.

[Claim 25]The device according to claim 24 whose query is a query of the full text.

[Claim 26]The device according to claim 24 with which a search system contains a statistical search engine.

[Claim 27]A client computer for acquiring a query from a user and displaying two or more documents in an output document group in order of said rank, As for said server, a processor answers a command memorized by memory including said processor and said memory, including further a server connected to a client computer via network connection, a query is acquired from a client computer -- the device according to claim 24 which gives said two or more documents in an output document group to a client computer in order of said rank.

[Claim 28]The device according to claim 27 containing a server in which plurality of a server is separate.

[Claim 29]The device according to claim 27 with which a search system contains a statistical search engine.

[Claim 30]The device according to claim 29 whose network connection is the Internet or an Internet connectivity.

[Claim 31]A search engine answers a query and a record memorized from a repository for one each of said two or more of the documents in an output document group is searched, A record includes information which pinpoints a place where one each of said the documents in an output document group may be found out, The device according to claim 30 which a processor accesses one each of said the documents from a server of relation for it, downloads it according to a command memorized by memory and information included in a record, and is included in an output document group.

[Claim 32]A client computer which has said processor and said memory, A server connected to a client computer via network connection is included further, The device according to claim 24 which said server realizes said search system, answers a query given by a client computer, and gives said output document group to a client computer.

[Claim 33]The device according to claim 32 with which a search system contains a statistical search engine.

[Claim 34]The device according to claim 33 whose network connection is the Internet or an Internet connectivity.

[Claim 35]A search engine answers a query and a record memorized from a repository for one each of said two or more of the documents in an output document group is searched, A processor answers a command memorized by memory and information included in a record

including information which pinpoints a place where one each of said the documents in an output document group may be found out as for a record, The device according to claim 34 which accesses one each of said the documents from a server of relation for it, downloads it, and is included in an output document group.
[Claim 36]The device according to claim 24 which a computer answers a command memorized by memory again, including further a computer which has said processor and said memory, and realizes said search system.
[Claim 37]The device according to claim 36 with which a search system contains a statistical search engine.
[Claim 38]The device comprising according to claim 5:
Each of said 1st logical form and said 2nd logical form includes a list and the 2nd list of the 1st, [correspondence of one or more logical form triads] A form of each stem of a word of two words to which said logical form triad in said 1st list and said logical form triad in said 2nd list were semantically related respectively in a logical form graph of correspondence in one phrase each of said document in a query.
A relation showing semantic relations between two words specified beforehand.

[Claim 39]Said list of the 2nd of logical form triads relevant to one of said the documents said list of the 1st of logical form triads relevant to a query, or in an output document group, respectively -- said query -- or the device according to claim 38 which contains further a paraphrase of words and phrases relevant to one of said the documents.
[Claim 40]Again a score for said one document A node word in the 2nd [for said one document] logical form, Frequency or semantic contents of said node word in said one document, The device according to claim 38 which is the function as which frequency of a specific logical form triad for frequency of a node word in said one document specified beforehand or semantic contents, and said one document or the length of said one document was determined beforehand.
[Claim 41]. Said function is identically [to at least one of the logical form triads relevant to a query] in agreement. The device according to claim 38 defined by the type of semantic relations relevant to it in dignity assigned to each logical form triad which is the sum total of dignity taken over a logical form triad relevant to each of two or more of said documents in an output document group, and is in agreement.
[Claim 42]a processor answers a command memorized by memory -- arbitrary things of a logical form triad relevant to a query judging whether it is in agreement with arbitrary things of a logical form triad relevant to arbitrary documents in an output document group, and, a triad in agreement relevant to said arbitrary documents is specified -- for one each of the documents in said output document group which has at least one related logical form triad in agreement,

Weighting is carried out to a logical form triad which is [in said one document each] in agreement using a weight numerical value specified beforehand with semantic relations relevant to each of said logical form triad in agreement, one or more dignity for said one document is formed -- a score for said one document is calculated as a function of said one or more dignity -- the device according to claim 41 which ranks one each of said the documents according to said the score, and specifies the order of a rank.

[Claim 43]The device according to claim 42 whose order of a rank is the order of what to a small thing has large dignity.

[Claim 44]The device according to claim 38 which presents a group by whom the 1st of said entry for said output document group which a processor answers a command memorized by memory and has the continuous highest rank of a document in said output document group was specified beforehand.

[Claim 45]The device according to claim 44 with which two or more documents in an output document group consist of a document in said output document group which has at least one related triad in agreement.

[Claim 46]The device comprising according to claim 45:

A form of each stem of a word of two words to which each of said 1st logical form triad and said 2nd logical form triad was semantically related in a logical form graph of correspondence in one phrase each of said document in a query, respectively.

A relation showing semantic relations between two words specified beforehand.

[Claim 47]The device according to claim 38 with which said logical form triad relevant to one of said the documents said logical form triad relevant to a query or in an output document group contains further a logical form triad containing one of superordinate words or synonyms of said word.

[Claim 48]The device according to claim 38 which is coincidence with said same coincidence between said arbitrary things of a logical form triad relevant to a query, and said arbitrary things of a logical form triad relevant to arbitrary documents in an output document group.

[Claim 49]The device according to claim 38 with which a repository contains a data set.

[Claim 50]The device according to claim 38 whose query is a query of the full text.

[Claim 51]The device according to claim 38 with which a search system contains a statistical search engine.

[Claim 52]A client computer for acquiring a query from a user and displaying two or more documents in an output document group in order of said rank, As for said server, a processor answers a command memorized by memory including said processor and said memory, including further a server connected to a client computer via network connection, a query is acquired from a client computer -- the device according to claim 38 which gives said two or

more documents in an output document group to a client computer in order of said rank.
[Claim 53]The device according to claim 52 with which a server contains two or more individual servers.
[Claim 54]The device according to claim 52 with which a search system contains a statistical search engine.
[Claim 55]The device according to claim 54 whose network connection is the Internet or an Internet connectivity.
[Claim 56]A search engine answers a query and for one each of said two or more of the documents in an output document group, Search a record memorized from a repository and a record includes information which pinpoints a place where one each of said the documents in an output document group may be found out, The device according to claim 55 which a processor answers a command memorized by memory and information included in a record, accesses one each of said the documents from a server of relation for it, downloads it, and is included in an output document group.
[Claim 57]A client computer which has said processor and said memory, A server connected to a client computer via network connection is included further, The device according to claim 38 which said server realizes said search system, answers a query given by a client computer, and gives said output document group to a client computer.
[Claim 58]The device according to claim 57 with which a search system contains a statistical search engine.
[Claim 59]The device according to claim 58 whose network connection is the Internet or an Internet connectivity.
[Claim 60]A search engine answers a query and for one each of said two or more of the documents in an output document group, Search a record memorized from a repository and a record includes information which pinpoints a place where one each of said the documents in an output document group may be found out, The device according to claim 59 which a processor answers a command memorized by memory and information included in a record, accesses one each of said the documents from a server of relation for it, downloads it, and is included in an output document group.
[Claim 61]The device according to claim 38 which a computer answers a command memorized by memory again, including further a computer which has said processor and said memory, and realizes said search system.
[Claim 62]The device according to claim 61 with which a search system contains a statistical search engine.
[Claim 63]A method characterized by comprising the following for using a document memorized in an information retrieval system for searching from a repository.
Said system answers a query, searches two or more documents relevant to the query

memorized, has them, and a search system for specifying an output document group said method, the 1st logical form shows semantic relations between words relevant to a query including a step which answers a query and produces the 1st logical form for it a document in an output document group -- each -- to one [another], The 2nd logical form shows semantic relations between words relevant to a phrase in said one document including a step which acquires the 2nd corresponding logical form, The 1st logical form of a query
A step which ranks two or more documents in an output document group as a function with the 2nd logical form for one each of two or more documents in an output document group defined beforehand, and specifies the order of a rank.
A step which gives two or more entries relevant to an output document group memorized as an output in order of said rank.

[Claim 64]A method according to claim 63 of each entry being one of the correspondences of a document in an output document group, or being a record relevant to one document of said correspondence.
[Claim 65]the inside of the 1st logical form for a query, and an output document group -- each -- a way according to claim 64 each with the 2nd logical form for another document is a list of a logical form graph, its subgraph, or logical form triads, respectively.
[Claim 66]said step to acquire -- a document in an output document group -- said -- each -- for one [another], or it reads the 2nd logical form of correspondence from a storage -- or -- the inside of an output document group -- said -- each -- a method according to claim 65 containing a step which generates the 2nd logical form of said correspondence by analyzing one another document.
[Claim 67]Said function generates a score based on a relation between each of said 1st logical form relevant to a query, and said 2nd logical form relevant to said one document defined beforehand for one of said the documents, A method according to claim 66 containing a step which said step to rank ranks an entry memorized according to a score relevant to each document in an output document group, and specifies the order of a rank.
[Claim 68]said 1st logical form relevant to a query, or said 2nd logical form relevant to one of said the documents in an output document group -- respectively -- said query -- or a method according to claim 67 of containing further a paraphrase of words and phrases relevant to one of said the documents.
[Claim 69]A method comprising according to claim 68:
Each of said 1st logical form and said 2nd logical form One or more logical form triads, Said logical form triad in said 1st list and said logical form triad in said 2nd list including a list and a list [the 2nd] of the 1st of correspondences respectively, A form of each stem of a word of two words which were semantically related to a logical form graph of correspondence in one

phrase each of said document in a query, respectively.
A relation showing semantic relations between two words specified beforehand.

[Claim 70]A method according to claim 67 of being coincidence with said same coincidence between arbitrary things of said 1st logical form relevant to a query, and arbitrary things of said 2nd logical form relevant to arbitrary documents in an output document group.
[Claim 71]A method comprising according to claim 70:
Each of said 1st logical form and said 2nd logical form One or more logical form triads, Said logical form triad in said 1st list and said logical form triad in said 2nd list including a list and a list [the 2nd] of the 1st of correspondences respectively, A form of each stem of a word of two words which were semantically related to a logical form graph of correspondence in one phrase each of said document in a query, respectively.
A relation showing semantic relations between two words specified beforehand.

[Claim 72]A way according to claim 67 a repository contains a data set.
[Claim 73]A way according to claim 67 a query is a query of the full text.
[Claim 74]A way according to claim 67 a search system contains a statistical search engine.
[Claim 75]A system characterized by comprising the following contains a client computer further, and said method is a client computer.
A step which acquires a query from a user.
A step which displays two or more documents in an output document group in order of said rank is included, in [including further a server by which a system is connected to a client computer via network connection] a server said method, A step which acquires a query from a client computer.
A step which gives said two or more documents in an output document group to a client computer in order of said rank.

[Claim 76]A way according to claim 75 a search system is a statistical search engine.
[Claim 77]A way according to claim 76 network connection is the Internet or an Internet connectivity.
[Claim 78]In a search engine, answer a query and a step which searches a record memorized from a repository for one each of said two or more of the documents in an output document group is included further, In a server including information to which a record pinpoints a place where one each of said the documents in an output document group is found out, and to obtain, A method according to claim 77 of answering information included in a record, accessing one each of said the documents from a server of relation for it, downloading it, and containing further a step included in an output document group.

[Claim 79]In [said server realizes said search system, including further a server which is connected to a client computer via a client computer and network connection as for a system, and] a server said method, A method according to claim 67 of containing further a step which answers a query given by a client computer and gives said output document group to a client computer.
[Claim 80]A way according to claim 79 a search system contains a statistical search engine.
[Claim 81]A way according to claim 80 network connection is the Internet or an Internet connectivity.
[Claim 82]In a search engine, answer a query and a step which searches a record memorized from a repository for one each of said two or more of the documents in an output document group is included further, In a client computer including information to which a record pinpoints a place where one each of said the documents in a set of an output document may be found out, A method according to claim 81 of answering information included in a record, accessing one each of said the documents from a server of relation for it, downloading it, and containing further a step included in an output document group.
[Claim 83]A way according to claim 67 a system contains further a step to which said method realizes said search system in a computer including a computer.
[Claim 84]A way according to claim 83 a search system contains a statistical search engine.
[Claim 85]Again a score for said one document A node word in the 2nd [for said one document] logical form, Frequency or semantic contents of said node word in said one document, A method according to claim 67 of being the function as which frequency of a specific logical form triad for frequency of a node word in said one document specified beforehand or semantic contents, and said one document or the length of said one document was determined beforehand.
[Claim 86]A way according to claim 85 a repository contains a data set.
[Claim 87]A way according to claim 85 a query is a query of the full text.
[Claim 88]A way according to claim 85 a search system contains a statistical search engine.
[Claim 89]In [including a client computer further] a client computer in a system said method, A step which acquires a query from a user, and a step which displays two or more documents in an output document group in order of said rank are included further, In [including further a server by which a system is connected to a client computer via network connection] a server said method, A method according to claim 85 of containing further a step which acquires a query from a client computer, and a step which gives said two or more documents in an output document group to a client computer in order of said rank.
[Claim 90]A way according to claim 89 a search system contains a statistical search engine.
[Claim 91]A way according to claim 90 network connection is the Internet or an Internet connectivity.

[Claim 92]In a search engine, answer a query and a step which searches a record memorized from a repository for one each of said two or more of the documents in an output document group is included further, In a server including information to which a record pinpoints a place where one each of said the documents in an output document group may be found out, A method according to claim 91 of answering information included in a record, accessing one each of said the documents from a server of relation for it, downloading it, and containing further a step included in an output document group.
[Claim 93]Including a server by which a system is connected to a client computer via a client computer and network connection, said server is realized and said search system said method, A method according to claim 85 of containing further a step which answers a query given by a client computer in a server, and gives said output document group to a client computer.
[Claim 94]A way according to claim 93 a search system contains a statistical search engine.
[Claim 95]A way according to claim 94 network connection is the Internet or an Internet connectivity.
[Claim 96]In a search engine, answer a query and a step which searches a record memorized from a repository for one each of said two or more of the documents in an output document group is included further, In a client computer including information to which a record pinpoints a place where one each of said the documents in an output document group may be found out, A method according to claim 95 of answering information included in a record, accessing one each of said the documents from a server of relation for it, downloading it, and containing further a step included in an output document group.
[Claim 97]A way according to claim 85 a system contains further a step to which said method realizes said search system in a computer including a computer.
[Claim 98]A way according to claim 97 a search system contains a statistical search engine.
[Claim 99]A method comprising according to claim 67:
Each of said 1st logical form and said 2nd logical form includes a list and the 2nd list of the 1st, [correspondence of one or more logical form triads] A form of each stem of a word of two words to which said logical form triad in said 1st list and said logical form triad in said 2nd list were semantically related respectively in a logical form graph of correspondence in one phrase each of said document in a query.
A relation showing semantic relations between two words specified beforehand.

[Claim 100]Said list of the 2nd of logical form triads relevant to one of said the documents said list of the 1st of logical form triads relevant to a query, or in an output document group, respectively -- said query -- or a method according to claim 99 of containing further a paraphrase of words and phrases relevant to one of said the documents.

[Claim 101]Again a score for said one document A node word in the 2nd [for said one document] logical form, Frequency or semantic contents of said node word in said one document, A method according to claim 99 of being the function as which frequency of a specific logical form triad for frequency of a node word in said one document specified beforehand or semantic contents, and said one document or the length of said one document was determined beforehand.
[Claim 102]. Said function is identically [to at least one of the logical form triads relevant to a query] in agreement. A method according to claim 99 defined by the type of semantic relations relevant to it in dignity assigned to each logical form triad which is the sum total of dignity taken over a logical form triad relevant to each of two or more of said documents in an output document group, and is in agreement.
[Claim 103]A method comprising according to claim 102:
Said step to rank, A step which judges whether arbitrary things of a logical form triad relevant to a query are in agreement with arbitrary things of a logical form triad relevant to arbitrary documents in an output document group, and specifies a triad in agreement relevant to said arbitrary documents
For one each of the documents in said output document group which has at least one related logical form triad in agreement, A step which carries out weighting to a logical form triad in agreement in said one document each using a weight numerical value specified beforehand, and forms one or more dignity for said one document with semantic relations relevant to each of said logical form triad in agreement.
A step which calculates a score for said one document as a function of said one or more dignity.
A step which ranks one each of said the documents according to said the score, and specifies the order of a rank.

[Claim 104]A way according to claim 103 the order of a rank is the order of what to a small thing has large dignity.
[Claim 105]A way according to claim 99 a step which gives an entry memorized contains a step which presents a group by whom the 1st of said entry for said output document group which has the continuous highest rank of a document in said output document group was specified beforehand.
[Claim 106]A way according to claim 105 said two or more documents in an output document group consist of a document in said output document group which has at least one related triad in agreement.
[Claim 107]A method comprising according to claim 106:
A form of each stem of a word of two words to which each of said 1st logical form triad and

said 2nd logical form triad was semantically related in a logical form graph of correspondence in one phrase each of said document in a query, respectively.
A relation showing semantic relations between two words specified beforehand.

[Claim 108]A way according to claim 99 said logical form triad relevant to one of said the documents said logical form triad relevant to a query or in an output document group contains further a logical form triad containing one of superordinate words or synonyms of said word.
[Claim 109]A method according to claim 99 of being coincidence with said same coincidence between said arbitrary things of a logical form triad relevant to a query, and said arbitrary things of a logical form triad relevant to arbitrary documents in an output document group.
[Claim 110]A way according to claim 99 a repository contains a data set.
[Claim 111]A way according to claim 99 a query is a query of the full text.
[Claim 112]A way according to claim 99 a search system contains a statistical search engine.
[Claim 113]In [including a client computer further] a client computer in said system said method, A step which acquires a query from a user, and a step which displays two or more documents in an output document group in order of said rank are included, In [including further a server by which said system is connected to a client computer via network connection] a server said method, A method according to claim 99 of containing further a step which acquires a query from a client computer, and a step which gives said two or more documents in an output document group to a client computer in order of said rank.
[Claim 114]A way according to claim 113 a search system contains a statistical search engine.
[Claim 115]A way according to claim 114 network connection is the Internet or an Internet connectivity.
[Claim 116]In a search engine, answer a query and for one each of said two or more of the documents in an output document group, In a server including information for which a record pinpoints a place where one each of said the documents in an output document group may be found out including further a step which searches a record memorized from a repository, A method according to claim 115 of answering information included in a record, accessing from a server relevant to [for it] one each of said the documents, downloading it, and containing further a step included in an output document group.
[Claim 117]Including a server by which a system is connected to a client computer via a client computer and network connection, said server is realized and said search system said method, A method according to claim 99 of containing further a step which answers a query given by a client computer in a server, and gives said output document group to a client computer.
[Claim 118]A way according to claim 117 a search system contains a statistical search engine.
[Claim 119]A way according to claim 118 network connection is the Internet or an Internet

connectivity.
[Claim 120]In a search engine, answer a query and for one each of said two or more of the documents in an output document group, In a client computer including information for which a record pinpoints a place where one each of said the documents in an output document group may be found out including further a step which searches a record memorized from a repository, A method according to claim 119 of answering information included in a record, accessing one each of said the documents from a server of relation for it, downloading it, and containing further a step included in an output document group.
[Claim 121]A way according to claim 99 a system contains further a step to which said method realizes said search system in a computer including a computer.
[Claim 122]A way according to claim 121 a search system contains a statistical search engine.
[Claim 123]A medium for memorizing a command which can be executed by computer and performing the step according to claim 63 which can be computer read.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]
This invention relates to the method of accompanying the device for an information retrieval system and it which improve overall accuracy using natural language processing, in order to process the result searched, for example with information retrieval engines, such as the conventional search engine based on statistics.
[0002]
[Explanation of the advanced technology]
In order to retrieve memory information from large capacity data storages, such as the conventional database containing a publication and/or the bibliographic information for it, automatic information retrieval technique is increasingly used frequently, until it results from tens of now. Such a conventional database is maintained by for example, U.S. electrical-and-electric-equipment Institute of Electronics and Communication Engineers (IEEE), For example, by the dialog (Dialog) information service of Knight-Ridder Information Inc. and like [now / in the case of an accessible INSPEC database], Although it is wide ranges, such as electrical engineering and a computer pertinent art, since the information turned to a specific topic is generally included, there is a tendency which becomes special (DIALOG is a registration service mark of Knight-Ridder Information Inc.). Although this type of database is sure clearly to continue increasing with the increase in the number of publication of the paper of relation, and other publications, this increase is comparatively gently-sloping and there is a tendency adjusted well moderately. The database specialized in this way is easy to be systematized comparatively well.
[0003]
however -- passing the Internet -- the so-called appearance of accessible "World Wide Web" (a

"web" is only called below) -- and more prosperous, In contrast with the conventional publication, contribute to a web, and information and the amount of information which can be used on a web by [which access information] being comparatively easy and being able to carry out by low cost, although not explosive, a near increase is accomplished very exponentially, and restriction is in a realistic field of view -- as -- it is not considered. **** by which a web attains to all the learning fields which human being tries about -- although a huge quantity of information is provided, information content on a web is not systematized, so that it is dramatically disorderly and going too far, therefore it complicates access and search of the information from a web dramatically, and forces it great labor.

[0004]

In order to make search of the information from a web easy far, it will be KOMPI in several [past]. The search engine of a large number by YUTA is developed, and, generally it is used widely. ******. Generally, these conventional engines were mounted by software. By a "web crawler." going into a website automatically -- the inside of it -- hyper--- pursuing - text link one by one, and extracting each document in it -- abstract carrying out and attaching an index to it by what is called a "keyword" -- large-sized data bay It is considered as SU and enables it to access behind. Specifically, it is such an abstract. Such each document that a crawler encounters is condensed by even what is usually called "one bag of word." This is all the information on semantics and syntax. Although removed, the content word which exists in a document is included. A content word is DOKI. It may be in YUMENTO itself and/or is the hyper of the document. It appears only in the description field of the - text markup language (HTML) version. There are also things. Anyway, ene [as opposed to such a document in an engine] Tori, i.e., a document record, is created. An index is attached to each of the content word about each document, and it has a link which returns to the document. ** and the data structure which can be searched are formed. Typically, a document record is (a) web address, i.e., URL. -- It is a web browser by this. Uniform resaw which can access the document of correspondence SUROKETA, (b) various content words in the document are included -- ** ** -- depending on an engine. this to other content words in the document ** -- a content word -- including each relative address, including the (c) outline furthermore -- this or it is only several lines in a document -- or the first several lines of a document although there are many a certain things -- further -- a case -- d -- the description field of the HTML The explanation about the document indicated inside is included. It is SA about a database. In order for - CHI to carry out, a user gives the query based on a keyword to an engine.

[0005]

case of 1 to which the query was typically given by the user or 2 or more, and many **** -- depending on engine capability including a small number of keywords, case **** -- the Boolean logic (for example, "AND" or "OR") between the continuous keywords or a similar operator (for

example, numerical nearness) is included. To a query Answer, and an engine is as logical as as many keywords as possible, or it is nearness. When the related operator is provided, Requested specific ****** Key which is in ** or the "range" (a specific number of content words) which exists mutually It is going to trace a document including a word. When performing this, an engine searchs the database and is in agreement with one of the keywords of a query. At least one word is included, When there is a request, it is for the request. DOKYU which is in agreement with the operator and/or range which were therefore specified MENTO is traced. each of such a document that found out the engine receiving and searching the document record about it -- inside of the document the same according to the number of keyword coincidence -- otherwise, I will go away -- the document carried out is received -- ranking It carries out and the record is shown to a user.

[0006]

With a query, only most documents which answered the query of the keyword given by the user and were searched are completely unrelated in many cases, and it irritates a user.

[0007]

Therefore, in order to reduce the number with which an unrelated document is searched, the conventional search engine (a "statistical search engine" is only called hereafter) based on a keyword has taken in the statistical procedure by those searching methods. For example, the total of the keyword which is in agreement between the keyword in a query, and the content word in each searched document record and the grade of coincidence of these words, That is, based on whether it is in the range of the nearness requested as combination, a statistical search engine calculates the numerical measure comprehensively called a "statistic" in many cases to each of such a searched document record. These statistics may include the reciprocal of the document frequency to each word in agreement. After that, an engine ranks a document record with those statistics, and returns a user a concerning 5 to 20 or less searched records of small number specified beforehand typically document record with the highest rank. The user to the 1st group's searched document is the 1st group's document record (or like an engine of a certain kind). When a document is returned with an engine and a user examines the document itself, a user, Next, the group of the document record of high ranking can be required, and it can require until all the document records searched like the following are examined in this way.

[0008]

Conventionally, the performance of a search engine has been estimated by reappearance and accuracy. Reappearance measures the number of such documents that answered the given query and were actually searched by percentage to all the documents of the relation in a data set. On the other hand, accuracy actually measures the number of the documents relevant to a query by percentage to all the searched documents. Since the mere number of the documents

searched eventually is not important, in considering a web search engine, it thinks that reappearance is not the important metrics about performance. Depending on a query, this number also actually has too large a thing. Therefore, in order to produce a useful result, I think that it is not necessary to take out all the documents of the relation by which the indexing was carried out with the engine. However, it is thought that accuracy is very important. That is, a rank thinks that the document which a user is shown first most highly should be a thing most relevant to a query.

[0009]

That is [the word of the thing whose accuracy of the conventional statistical search engine is comparatively low is an independent variable], it originates in the word of all the texts being based on assumption of appearing independently mutually. The conditional probability that one arbitrary word appears in that document when a certain word different from the independence in this case exists in a document is always zero, That is, a document only contains the meeting of a word without structure, or means being only "one bag of only word." This assumption is dramatically mistaken about all the languages so that it can recognize easily. In a huge quantity about a word, have English like other languages and the structure on complicated syntax and lexical-semantics the meaning of these words, Based on the specific linguistic context used, it is often large, and differs in many cases, and what kind of word determines whether it appears behind as the meaning in which the context was given to the word also in such a case. Therefore, if the word which appears in a text is only independently, there is, and it depends mutually highly. [no] The search engine based on a keyword has completely disregarded a linguistic structure fine [this]. For example, "How many hearts does an octopus have?" expressed with natural language

The example of the query to say is considered. If it is a statistical search engine which operates based on a content word "hearts" and "octopus", or its morphological-like stem of a word, As the word which expresses the contents according to the portion of the material "squid (cuttlefish) from artichoke hearts (core of arch chalk), onions (onion), and octopus (octopus) " -- the memorized document containing the recipe which it has is returned, or a user is led to the document. Since two contents words "octopus" and "hearts" showing the contents are in agreement, this engine, For example, based on the statistical measurement value containing nearness and a logical operator, even if a document is quite unrelated actually to a query, it will determine to be the coincidence excellent in this document.

[0010]

In this technical field, various policies for extracting the element of the syntax phrase as head correction **** which has a relation which it is not with a label are taught. The indexing of these elements is carried out after that as a predicate in the conventional statistical vector space model (there is no internal structure typically).

[0011]
An example of such a policy is J. L. Fagan,"Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-. It is indicated to Syntactic Methods", Ph.D. Thesis, Cornell University, 1988, and pages i-261. This policy analyzes an English text using natural language processing, the component of a syntax phrase is specifically extracted, the component of these phrases is behind treated as a technical term, and an indexing is carried out to the index using a statistical vector space model. At the time of search, a ** user inputs a query with natural language, under this policy, natural language processing is performed and analyzed by this query, and the element of the component of a syntax phrase similar to the feared element by which the record is carried out to search among an index is extracted. then, the **** trial with which comparative collation coincidence ** is performed in the component of the syntax phrase by a query and ** memorized by the index is made. The author is making the statistical policy contrast, if the stochastic method is used, since [this] the component of a syntax phrase is purely specified for a syntax policy.
By this statistical policy, since the component of a syntax phrase is specified, the stochastic method is used and it is.
an author -- the accuracy with a ** stochastic policy to ** large in natural language processing which does not find and carry out an improvement but natural language processing can sometimes provide substantially -- ** -- in what justifies whether it follows on natural language processing depending on the slightly minor thing improved, and the becoming cleanup cost, it is concluded that it cannot do.
[0012]
The case where the natural language processing for choosing the suitable word for making it contain in the query for a search is used, Another policy based on such syntax, T. Strzalkowski and "Natural. Language Information. Retrieval: TIPSTER-2. Final Report" and Proceedings. of Advances in Text ProcessingTipster Program Phase 2, DARPA, and 6-8 May 1996, Tysons Corner, Virginia, pages 143-148 (hereafter) : T., and it calls it "the paper of DARPA." Strzalkowski and "Natural. It is indicated to Language Information Retrieval", Information Processing and Management, Vol. 31, No. 3, 1995, and pages 397-417. although this policy is what had possibility and is stretched theoretically, since the processing demanded in order that an author may realize natural language art used as a base in 148 pages from the 147th page of the paper of DARPA is an altitude, it has been concluded that this policy of the present place is not practical.
"In the capability to treat the sentence of natural language, ..., however the NLP [natural language processing] art of satisfying our performance requirements (or it is thought that it is close to these requirements at least) still have quite low performance. In particular, processing in connection with notional composition, a logical form, etc. which progressed is not still

reached from the field of calculation. since such art which progressed is the things coping with the problem of the limit of an expression level, it becomes more effective -- although it can also assume that it will come out, it is deficient in an actual proof and must be limited to the test of a moreover quite small scale. "

The further policy based on this kind of integrated theory B. Katz and "Annotating the. World Wide Web Using. Natural Language" and Conference. Proceedings of RIAO. 97 and Computer-Assisted. Information Searching in Internet, McGill University, Quebec, Canada, 25-27 June 1997, Vol. 1, pages 136-155 (hereafter) It is indicated to call "the publication of Katz." Since subject-verb-object expression is created according to the statement of the publication of Katz, with an internal structure maintained, the change on minor syntax can be coped with at the time of search.

[0013]

That these syntactic-like policies did not bring about a great improvement or by the ability to have not realized depending on the available natural language processing system at the time, an area of research, From the trial which it is going to improve directly, the accuracy as a result of the beginning by a query, and reappearance to the improvement of a user interface, i.e., a concrete target. It shifted to the improvement by the method for visualizing the result which receives a query including dividing into the method where "searching a class word" and a user raise the accuracy of a query based on an interaction with users, such as answering search results, and a suitable lump, and displaying a result.

[0014]

Although these improvement itself is useful, there is so little improvement in the accuracy which can be attained by these improvements that it is still disappointed, and it is clearly insufficient against reducing substantially the impatience which a user peculiar to the search by a keyword feels. Specifically, a user is still required to screen a set of a comparatively big document which a related response does not spread sparsely by manual operation.

[0015]

Therefore, in this technical field, the art for retrieving information and the method of accompanying a device and it specifically of bringing about an improvement of the remarkable accuracy which surpasses what can be attained by the conventional statistical policy to information retrieval are needed. Such art needs to bring about the result which it can be reliable and it can repeat also to the type and length of a wide range sentence in the text produced arbitrarily, and needs to be practical, and needs to be effective also at the cost aspect in the case of realization. In order to improve accuracy remarkably in spite of a problem peculiar to this technical field moreover to the accuracy of such a conventional policy, Such art should use natural language processing, in order to acquire the effect of choosing and searching the document of relation based on collation with semantic contents and the contents

of the query, and showing a user after that preferably.

[0016]

[Summary of the Invention]

According to our large instruction, by this invention, in order to improve the accuracy of the document search based on the keyword performed, for example by a statistical web search engine, natural language processing is adopted.

Therefore, this necessity is satisfied.

[0017]

Saying roughly, each generates and compares a search query and the logical form relevant to searched each of the document, and this processing includes carrying out weighting of those coincidence. Based on the function as which the query and the "logical form" about both searched documents were specified beforehand, specifically, based on the sum of the dignity relevant to the logical form in agreement relevant to a document, the searched document is ranked and is eventually displayed in the order. A logical form is a directed acyclic graph with which the word showing the text of arbitrary sizes is linked by the relation with a label. Especially a logical form draws especially the relation between the theme and a modifier for the relation on the semantics between the important words in an input string. This depiction can take various specific forms, for example, can take the form of a logical form graph or its arbitrary subgraph which includes the list of logical form triads (triple), for example. Although each of a triad takes the form of a "word-relative-word" here, either of these forms can be used for our invention.

[0018]

According to our specific instruction, the group of the document eventually searched by such search, for example from a database or World Wide Web is obtained. After that, natural language processing is performed to each document, processing about form [be / morpheme / and / syntax / it / logical] is specifically performed to it, and a suitable logical form is eventually generated to each sentence of each document. The query given by the user is analyzed similarly and a set of the logical form triad of the correspondence about it is obtained. The group of the logical form to a query is compared with the group of the logical form relevant to each of the searched document after that, and coincidence with the logical form from the group of a query and the logical form from the group of each document is checked. The document which does not produce coincidence is not taken into consideration any more. Score attachment of each remaining documents is carried out after that at heuristics. The dignity which, namely, was especially specified to a deep subject, a depths object, a functional word, etc. which may appear in a logical form beforehand is assigned. [the relation of various types] Each score of such remaining documents is the function as which the dignity of the logical form

in agreement in it was specified beforehand. The sum of the meaning dignity relevant to all triads (double coincidence is disregarded) in agreement which appears in that document may be sufficient as this function, for example. At the end, the held document is a group of the small number specified beforehand who tells five pieces or ten pieces typically to a descending order according to a user's selection based on those scores, It begins from the group who has the highest score, and a user is shown in the turn of other groups at order.

[0019]

This invention can be used by some various processing topology. Namely, when both searches (document search) based on the (a) query and a keyword are processed by common computers, such as a local personal computer (PC), (b) The search based on a keyword is processed by the remote computer which is a remote server, it comes out, when for example, the (c) query is created for a query and a search result with client PC and the remaining processing is distributed to various remote servers. [when processed by client PC] By pretreating, when the indexing of each document of a database is carried out and it is put in a database, obtaining the logical form of relation, memorizing these logical forms, and enabling it to access behind, When the document is searched behind and receives natural language processing, execution time always comes to be saved.

[0020]

[Detailed Description of the Invention]

If the following detailed explanation is taken into consideration with reference to an attached drawing, he can understand instruction of this invention easily.

[0021]

In being possible, it gives the same reference number to the element common to a drawing, in order to understand easily.

[0022]

If the following explanation is taken into consideration and it will be a person skilled in the art, regardless of whether a search engine is the conventional statistical engine, instruction of our this invention will be easily used for almost all information retrieval systems, It could recognize clearly that the accuracy of the search engine used by the system can be raised. When our invention retrieves the information on text format from almost all the types of the database etc. which are memorized by magnetic media, an optical medium (for example, CD-ROM), or other media of mass storage device, For example, irrespective of which language the information on text format is, English, Spanish, German, etc. can be used in order to improve accuracy.

[0023]

If it says widely, according to our this invention, the record provided by the search engine currently used, for example by the search engine, For example, ultimately, since a document is filtered and ranked, that it can be made to improve remarkably has recognized the accuracy of

a search engine to us by using natural language processing.

[0024]

With careful attention to this point, drawing 1 shows the block diagram of the very high level of the information retrieval system 5 using our invention. The system 5 contains the conventional search engine 20 which is a statistical search engine based on a keyword, and the processor 30 which continues after that. The processor 30 uses the natural-language-processing art of an invention of us so that it may explain later, The document generated with the engine 20 is filtered, it ranks again, and a set of the searched document higher than the case where the relevance over the query given by the user does not carry out specification of this art which ordering was made is brought about.

[0025]

Specifically, a user gives the query for a search to the system 5 at the time of operation. It makes the most of the contents on the semantics by natural language processing, and in order to raise accuracy further rather than the accuracy obtained by it only with the engine 20, full text (usually called "literal") form is used for a query. The system 5 gives this query to both the engine 20 and the processor 30. A set of the document which the query was answered, and the engine 20 searched the data set 10 of the memorized document, and was searched is outputted. A set (called an "output document group".) of this document is given to the processor 30 as an input so that it may be expressed with the line 25. Within the processor 30, to each of the document in a set, processing about natural language processing, especially form [be / morpheme / and / syntax / it / logical] is performed, and the logical form to each sentence in the document is generated so that it may explain in detail later. Each of such a logical form about a certain sentence codes the semantic relation especially the theme, and modifier structure between the words in the linguistic phrase in the sentence. The processor 30 analyzes a query in the same mode, and generates a set of the logical form of the correspondence about it. And the processor 30 compares a set of the form about a query with a set of the logical form relevant to each of the document in the set, and checks whether coincidence is between the logical form in a query set, and the logical form about each document. The document which does not induce coincidence is not taken into consideration any more. Each of the document containing at least one logical form which is in agreement with the logical form about the remaining queries is held by the processor 30, and score attachment is carried out at heuristics. The dignity specified beforehand is assigned to a deep subject, a depths object, a functional word, etc. which may appear, the relation, i.e., the logical form triad, of each different type, so that it may explain later. The dignity (namely, score) of each sum total of such a document is the sum of all the dignity of what disregarded the meaning triad which is in agreement, for example, i.e., the triad which is doubly in agreement. finally the processor 30 ranks the held document based on those scores -- for example, five

pieces and ten pieces -- it divides into a number to say of groups defined beforehand, and a user is shown from what has the highest score.

[0026]

The system 5 is dramatically general-purpose, and since wide range various applications can be made to suit, in order to simplify the following arguments, we decide to use one example and to argue about the use of an invention of us. This example is a document from World Wide Web, and probably, it is an information retrieval system which adopts the statistical Internet search engine based on the conventional keyword, in order to search the record in which the document of English which forms the data set by which the indexing was carried out was stored. Generally, such each record includes the information about the document of correspondence specified beforehand so that it may explain below. In the case of other search engines, a record may also contain the whole document itself. Although our invention is treated by making into an example the case where it is used for the conventional Internet search engine which searches with the following arguments a record including a certain information including the web address about a document which can find out the document, Speaking comprehensively, the ultimate item searched with the engine being a document actually, though the interim processing using the address is generally adopted in order to actually access a document from a web. Consideration of the following explanation will permit easily for our this invention to be able to suit easily use of the example of application of any of other information retrieval to the person skilled in the art.

[0027]

Drawing 2 shows the block diagram in the high level of the specific example of the invention of us currently used in the example of the Internet search engine. Our invention mainly describes this specific example in detail as an example. Including the computer system 300 of a client personal computer (PC) etc., the system 200 passes the network connection 205 and this is the network 210 (although it is the Internet here) so that it may be illustrated. For example, and it may replace such other networks, such as intranet, with this and may use them, it is connected to the server 220 by the network connection 215. As for a server, this acts as a host of the Internet search engine 225 including the computer 222 typically, For example, an ALTAVISTA search engine (ALTA VISTA is a registered trademark of Digital Equipment Corporation of State Maynard of Massachusetts (Maynard, Massachusetts).) is a type, it is connected to the large capacity data storage 227, and, typically, this is a data set of an accessible document record by the World Wide Web on index attachment ** and the Internet in a search engine. Each of such a record A web address (usually referred to as uniform resource locator URL) typically accessible in the document of correspondence by the (a) web browser, (b) Including the content word which appears in the document and which was specified beforehand depending on an engine. The outline which is only several lines in the (c)

document further including the relative address of such each word to other content words in the document, or is the first several lines of a document in many cases, (d) Include explanation of a document with which the HyperText-Markup-Language (HTML) description field is provided.

[0028]

The user stationed at the computer system 300, The web browser of relation which operates by this system (for example, from Microsoft Corporation, it is available and) The Internet connectivity to the server 220 and the search engine 222 which operates especially there is established via the thing based on the "Internet Explorer" version 3.0 which changed suitably so that instruction of an invention of us might be included. Furthermore, a user inputs into a browser the query expressed by the line 201 here, and a browser transmits a query to the search engine 225 by the Internet connectivity to the server 220 via the system 300. Then, a search engine processes a query to the document record memorized by the data set 227, and generates the set of a retrieval record to the document judged that an engine relates to a query. The engine 225. [how the indexing of the document is carried out, a document record is formed, and it memorizes to the data storage equipment 227, and] And in order to choose such a memorized document, since it is unrelated to this invention, what kind of analysis an engine conducts actually explains no these aspects of affairs [each of] to details more. It is enough if it says that a query is answered and the engine 225 returns a set of the searched document record to the web browser 420 via an Internet connectivity. At the same time as the engine 225 searches a document, after that, the browser 420 analyzes a query/or after that, and generates a set of correspondence of the logical form triad. If a search engine completes the search, a set of a document record is taken out and the set is given to a browser, The document (namely, thing which forms a set of an output document) of correspondence itself is accessed by the browser from the web server of relation (the data set relevant to this forms the "repository" of the document saved as a whole.). Also in the data retrieval application etc. of the CD-ROM base which became independent, for example, such a repository may be a stand-alone data set like. Then, a browser analyzes each of the accessed document (namely, thing of the output documents), and forms the set to which the logical form triad about each of such a document corresponds. Then, so that it may explain in detail later the browser 420, Based on coincidence collation of the logical form triad between a query and the searched document, Carry out score attachment of each document which has such coincidence, and those documents, According to selection of the user who was in secret touch with the descending order of the rank in the browser so that it might be expressed by the line 203, as a group of a document by whom a small number which has the typical highest rank was specified beforehand, The next group continues and it is the same as that of the following until it is ranked from the thing of a big score, a user is shown and a user checks further the

sufficient number of the documents shown in this way after this. In order that drawing 2 may gain a document record and a document from a remote server, our invention is shown as what uses network connection in illustration, but our invention is not limited in this way. So that it may explain in detail below in relation to drawing 9 A Retrieval application and a computer with our common invention, That is, it performs on local PC and the accompanying data set which was memorized by CD-ROM or other suitable media there is arranged, and if accessible, such network connection is unnecessary.

[0029]

Drawing 3 shows the block diagram of the computer system 300 shown in drawing 2, and this computer system 300 takes in instruction of ********.

[0030]

This system that is a client personal computer so that it may be shown, The input interface (INPUT I/F) 330 by which interconnection is altogether carried out by bus 370 from the former, the processor 340, the communication interface (COMM I/F) 350, the memory 375, and the output interface (OUTPUT I/F) 360 are included. Although the memory 375 generally contains various forms (in particular these all do not show for simplification) which are random access memory (RAM) and hard disk memory storage, The operating system (O/S) 378 and the application program 400 are memorized. The software which mounts instruction of an invention of us is typically built into the application program 400, and it is included in a web browser (shown in drawing 4) especially in this example. This operating system, The Windows NT operating system (it is available now from Microsoft Corporation (the registered trademark "Windows NT" is owned) of Redmond, Washington (Redmond, Washington).) etc., It may be mounted by what kind of conventional operating system. Since the process which is a component of O/S 378 is unrelated, it does not explain each portion to be an invention. However, a browser, therefore the software of an invention of us are also incorporable into the operating system itself. However, it is assumed that our browser is disengageable from an operating system because of the purpose of illustration and simplification, and it is in the application program 400. The application program 400 is executed under control of O/S 378. As opposed to each of the execution application program containing a web browser, Each command which the user specified in the instance of one or two separate tasks or more, When selectable commands, such as a menu and an icon in a tool bar, are typically available, and a user operates the vice 390 appropriately in an input, the command inputted interactively is answered, it is called by the user, and the display 380 is shown the accompanying information.

[0031]

As shown in drawing 3, the information which carries out ingress is produced, for example from two external sauce. Namely, from other equipment (on the whole, all are shown in drawing 2

as the network 210) with which the Internet, intranet, etc. were connected by network, for example. It reaches from an input source for exclusive use to the input interface 330 via the information which is attained to the communication interface 350 (shown in drawing 3) via the network connection 205 and by which network sky supply is carried out, or the course 310. an input for exclusive use is remote, for example as it is local -- or it is alike in their being other input sources, and it is not concerned but is generated from various sauce, such as an external data set. It is connected to the course 310 and the input interface 330 includes the suitable circuitry which provides the electrical connection of correspondence required in order to connect physically each of various sauce of input for exclusive use to the computer system 300 and to interface. The application program 400 under control of an operating system, Sauce, etc. and a command for exclusive use and data are exchanged via external sauce, such as a remote web server, or the course 310 via the network connection 205, and the information typically requested by the user at the time of execution of a program is transmitted and received.

[0032]

With the lead 395, the input interface 330 electrically connects the user input devices 390, such as a keyboard and a mouse, to the computer system 300, and interfaces. The printers 385, such as the display 380 of the conventional color monitor etc. and the conventional laser beam printer, are connected to the output interface 360 by the leads 363 and 367, respectively. In order that an output interface may electrically connect with a computer system and may make a display and a printer interface with it, it provides indispensable circuitry. The hard copy output information from the application under execution is given to a user with the printer 385. Especially the user stationed at the system 300 by operating appropriately a display, a printer, and the input device 390 (specifically a mouse and a keyboard), For example, via the Internet, it communicates using either of a huge number of the remote web servers which contain a still more nearly accessible search engine from there, and a screen, and in order to display and print locally, information, including a document etc., can be pulled out from there.

[0033]

Since each aspect of affairs of the software memorized by the specific hardware components and the memories 375 of the computer system 300 other than a thing required for realization of this invention is a thing from the former and is common knowledge, it is not explained to details any more.

[0034]

Drawing 4 shows the block diagram of the very high level of the application program 400 executed within the computer 300 shown in drawing 3. As these programs are shown in drawing 4 in the range relevant to this invention, this web browser 420 includes the retrieval

process 600 (in relation to drawing 6 A and drawing 6 B, explained in detail later) for realizing our this invention including the web browser 420. If it assumes that the Internet connectivity is established between a web browser and the statistical search engine which users, such as an ALTA VISTA search engine, chose, A user gives the query for a full text ("literal") search to the process 600 so that it may be expressed with the line 422 shown in drawing 4. This process transmits a query to a search engine via a web browser so that it may be expressed with the line 426. Although not shown in particular, the process 600 analyzes a query inside further, and generates the logical form triad of the correspondence, and these are later memorized locally in the computer 300. Answering a query, a search engine gives a set of the document record statistically searched so that it might be expressed with the line 432 to the process 600. On the web address and specific target which can access the document of ** as above-mentioned, each of these records URL, Sufficient suitable command to download the computer filing which is demanded by a remote web server with the document and containing the document via the Internet is included. When the process 600 receives all records, this process, Via the web browser 420, a suitable command tends to be transmitted so that it may be expressed by the line 436, all the documents (namely, thing which forms a set of an output document) specified with the record tend to be accessed, and it is going to download them. And sequential access of these documents is carried out from the web server corresponding to them, and they are specifically downloaded in the process 600 at the web browser 420 so that it may be expressed with the line 442. If these documents download, the process 600 will analyze each of such a document, will generate the logical form triad of the correspondence about it, and will memorize it locally. Then, by comparing the logical form triad about a query to the logical form triad about each document, So that the process 600 may carry out score attachment of each document containing at least one logical form triad in agreement, may rank these specific documents based on those scores and may finally be expressed by the line 446, It is directed in descending order of the score of a document that these specific documents show a user to the browser 400 for every group. The browser 400 creates a selection button suitable on the screen of the display 380 (refer to drawing 3), and by this a user, It can choose by "clicking" it appropriately with a user's mouse, and each of the group of the document which follows can be displayed according to a request.

[0035]

In order to judge the information on semantics, to save and to fully evaluate the usefulness of the logical form at the time of coding, At this time, it separates, the logical form and logical form triad which are used for this invention are illustrated and explained in the related range from explanation of the processing which realizes our invention, and the mode by which they are generated is explained briefly.

[0036]

Saying roughly, a logical form is a directed acyclic graph with which the word showing the text which has arbitrary sizes is linked by the relation by which label attachment was carried out. Although a logical form may include a superordinate word and/or its synonym in the important word in a phrase, and this word, it draws the semantic relation between these words. The logical form can take either among many various gestalten, for example, can take the form of a logical form graph or its arbitrary subgraphs, such as a list of logical form triads, so that it may be explained and illustrated with reference to drawing 5 D from drawing 5 A. For example, each of these triads has form of a "word-relation child-word." Although this invention generates and compares a logical form triad in this example, this invention can use easily what kind of form of others which were mentioned above, if the semantic relation between words can be drawn.

[0037]

Since he can understand logical form triads and those structures best by making into an example a series of sentences which become complicated by order, refer to the drawing 5 A for them first. This figure shows the logical form graph 515 and the logical form triad 525 about an illustration input string and the sentence specifically "The octopus has three hearts."

[0038]

Generally, in order to generate the logical form triad to the input string 510, for example, an input string, purging of this character string is carried out first, and it is decomposed into the word of that component. Then, the record (don't get confused with the document record adopted by a search engine.) in the dictionary stored beforehand defined beforehand is used to such each word, The record of correspondence is combined with the word of these components by the grammatical rule made beforehand, and it becomes a big structure or syntax, and further, they are again combined by the grammatical rule defined beforehand, and form a still bigger structure of an syntax-analysis tree etc. by it. A logical form graph is built from an analysis tree after that. Selectively, it is governed by whether the attributes and those values of a certain correspondence exist in a word record whether a specific rule can apply to a set of a specific component. And this logical form graph is changed into a series of logical form triads. For example, our invention uses the dictionary which has an entry of about 165,000 head words. This dictionary contains words of various classes, such as the preposition, the conjunction, the verb, noun and operator which specify the characteristic on syntax peculiar to the word in an input string, and semantics, and a quantifier, so that the analysis tree about an input string can be built. Clearly, it is a logical form (or about the problem). The logical form triad or logical form graph in a logical form which can draw the relation on semantics is not calculated when the document of correspondence behind is searched, While the indexing of the document of correspondence is carried out, it calculates beforehand, for example, stores in the record about the document, and can be used for next

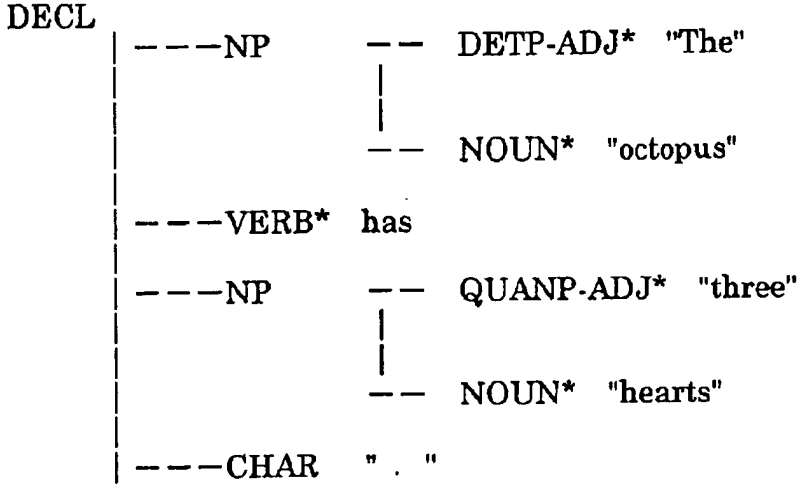
access. So that another example of an invention of us who explain later in relation to drawing 13 B from drawing 10 may see, Thus, it calculates beforehand, and if stored, the execution time of relation required in order to treat the document which it is effective in the quantity of natural language processing decreasing dramatically, therefore was searched according to our invention will become short.

[0039]

Morphological analysis is first conducted using the record which is in a dictionary about each of the word of the component and which was specified beforehand, and especially input strings, such as the sentence 510 shown in drawing 5 A, generate what is called "stem-of-a-word" (or "base") form about it. A stem of a word is used in order to normalize the form of various words, such as modification of a noun called verbal tense and singular number-plurality, in the common morphological-like form which a purser can use, for example. Once stem-of-a-word form is created, using the attribute in a grammatical rule and the record of the word of a component, the syntax of an input string will be analyzed by the purser and the syntax-analysis tree about it will be generated. This tree is shown and the structure of an input string specifically, For example, each word or phrase like the noun phrase "The octopus" in an input string, a category of the grammatical function of correspondence, for example like NP to a noun phrase, and the link to each word or a phrase related syntactically in it are expressed. It will be as [sentence / illustration / 510] follows [tree / of relation / syntax-analysis].

[0040]

[Table 1]



[0041]

Table 1 The start node in the upper left side of the syntax analysis tree tree about -- "The octopus has three hearts." defines the type of an input string by which purging is carried out. "QUES" about "IMPR" about "DECL" (upper example) about a declarative sentence and an imperative sentence and an interrogative sentence is contained in the type of a sentence. The

syntax of the 1st level is displayed on the lower right perpendicular of a start node. The head node in which this syntax is a main verb (the word in this example "has"), typically and which was shown by the asterisk, (In this example, it is the noun phrase "The octopus") It has an introduction modifier and a modifier which comes after that (it is the noun phrase "three hearts"). Each of leaves contains the word or punctuation contained in a dictionary. Here, "NP" shows a noun phrase as a label and "CHAR" shows a punctuation.

[0042]

And an syntax-analysis tree is further processed using the rule of a different group, and logical form graphs, such as the graph 515 about the input string 510, are generated. Including the process of creating a logical form graph extracting a lower layer structure from syntax analysis of an input string, while a logical form graph is mutual, it contains two or more words defined as having the functional character of semantic relation and its relation. The following is contained, the rank (Case), i.e., the functional role, of the "depths" which are used for the classification of various semantic relation.

[0043]

Dsub -- Deep subject Dind -- Depths indirect object Dobj -- Depths object Dnom -- Depths predicate nominative case Dcmp -- Depths objective complement Table 2 Since all the semantic relation in an input string is specified, each node of the syntax-analysis tree about the character string is inspected. In addition to the above-mentioned relation, other following semantic roles are used, for example.

[0044]

PRED -- Predicate PTCL -- Particle Ops in the verb of two-copy composition -- Functional word, For example, number Nadj -- Adjective which embellishes a noun Dadj -- Predicative adjective PROPS -- Ornamentation part which is a paragraph and which is not specified as others MODS -- Ornamentation part which is not a paragraph and which is not specified as others Table 3 An additional semantic label is specified as follows in a similar manner.

[0045]

TmeAt -- Time LocAt -- Place Table 4 The result of such analysis of the input string 510 is the logical form graph 515 anyway. Mutually, an eclipse with a link and the relation between them are specified as a linking attribute (for example, Dsub), and the word in which semantic relation (for example, between "Octopus" and "Have(s)") is accepted while it is mutual with the word in an input string is shown. This graph has caught the structure of the theme about each input string, and a modifier so that it may be represented by the graph 515 about the input string 510. Logical form analysis especially maps functional words, such as a preposition and an article, in the relation on the feature shown in the graph, or structure. Logical form analysis detects solution, i.e., the suitable functional relationship specify the right precedence relation between a pronoun and a coreference noun phrase, for example, and concerning an

abbreviation further, for an anaphora further, and is shown. At the time of logical form analysis, in order to cope with ambiguity and/or other linguistic idiosyncrasy, processing may be performed further. And in the conventional mode, the logical form triad of correspondence is read from a logical form graph, and is memorized as a group. As shown in a graph, three each group contains [it is mutual] the node word of two ******** with a link with semantic relation, between. About the input string 510 as an example, the logical form triad 525 is obtained from the processing graph 515 as a result. Here, the logical form triad 525 contains three separate triads which give semantic information peculiar to the input string 510 as a whole.

[0046]

Similarly as shown in drawing 5 D from drawing 5 B, the input strings 530, 550, and 570, i.e., the sentence of illustration, "The octopus has three hearts and two lungs.", "The octopus has three hearts and it canswim.", And to "I like shark fin soup bowls.", the logical form graphs 535, 555, and 575 and the logical form triads 540, 560, and 580 are obtained as a result, respectively.

[0047]

Apart from the thing of the conventional mode in which a logical form triad contains conventional "graph WOKU" generated from a logical form graph, there are three logical form structures which need the natural language processing of an addition required in order to obtain all the logical form triads correctly. In the case of an equistasis word [as / in the sentence 530 of illustration "The octopus has three hearts andtwo lungs.", i.e., an input string,], the logical form triad to a word, its semantic relation, and each of the value of the component by which equistasis was carried out is generated. According to "special" graph WOKU, it turns out that two logical form triads "have-Dobj-heart" and "have-Dobj-lung" are shown in Drawing 540. Probably, only one logical form triad "have-Dobj-and" was obtained when only conventional graph WOKU was used. Similarly in the case of the component which it has, a reference term (Refs) The sentence of illustration "The octopus has three hearts and it canswim.", That is, the logical form triad to a word, its semantic relation, and each of the value of a Refs attribute is generated besides the triad generated by conventional graph WOKU like [in the case of the input string 550]. According to this special graph WOKU, the logical form triad "swim-Dsub-octopus" other than the conventional logical form triad "swim-Dsub-it" is found out by the triad 560. When modifiers of a noun are consisted of as in the sentence 570 of illustration "I like shark fin soup bowls.", i.e., an input string, by the last, since the possible internal structure of the compound of a noun is expressed, the further logical form triad is generated. In conventional graph WOKU, the logical form triad "bowl-Mods-shark", "bowl-Mods-fin", and "bowl-Mods-soup" reflecting a possible internal structure [[shark] [fin] [soup] bowl] are generated. In special graph WOKU, the following possible internal structure [[shark fin] [soup] bowl], [[shark] [fin soup] bowl] And the logical form triad of respectively an addition

since [[shark [fin] soup] bowl] is expressed "fin-Mods-shark", "soup-Mods-fin" and "soup-Mods-shark" are generated.

[0048]

Since specific syntax morpheme details of logical form processing are not related to this invention, still more detailed explanation is given to omit. However, the further details about this, It applies on June 28, 1996, The sequence number 08th / No. 674,610 were given. "the method and system ("Method and System for Computing Semantic Logical.) for calculating a semantic logical form from an syntax-analysis tree The U.S. patent application under simultaneous pendency entitled Forms from Syntax Trees"", Please refer to "the information retrieval ("Information Retrieval Utilizing Semantic Representation of Text") using a semantic expression of a text" for which it applied especially on March 7, 1997 and to which the sequence number was given.Each of these is transferred to the grantee of this application, and is used here by quotation.

[0049]

Suppose that it returns to the argument on processing which realizes our this invention bearing in mind the outlines and those structures of this logical form.

[0050]

The flow chart of the retrieval process 600 of the invention of us currently used for the specific example of the invention of us shown in drawing 2, drawing 3, and drawing 4 is packed into drawing 6 A and drawing 6 B, and is shown, and the exact arrangement of these drawings is shown in drawing 6. The remaining operation shown in these drawings other than the operation shown in the block 225 drawn with the dashed line is performed by the computer system which is client PC300 (refer to drawing 2 and drawing 3), for example, and is specifically performed within the web browser 420. In order to understand easily, drawing 2, drawing 3, and drawing 6 A and drawing 6 B should be referred to simultaneously, reading the following explanation.

[0051]

If it goes into the process 600, executive operation will progress to the block 605 first. If this block is performed, it will demand inputting the query of the full text (literal) into the web browser 420 from a user. A query a single question (for example, "Are there any air-conditioned hotels in Bali?"), A single sentence (for example, "Give me contact information for all fireworks held in Seattle during the month of July."), They may be a part of (for example, "Clothes in Ecuador") forms of a sentence. If this question is obtained, executive operation will branch for the course 645 via the block 610 and the course 643 via the course 607. If the block 645 is performed, it will call the NLP routine 700, will analyze a query, will build a set of the logical form triad of the correspondence, and will memorize it locally. If the block 610 is performed, it will transmit the query of the full text to remote search engines, such as the

engine 225 put on the server 220, from the web browser 620 by an Internet connectivity so that it may be expressed with the dashed line 615. At this time, the block 625 is performed by a search engine, answers a query, and takes out a set of a document record. If this set is formed, that set will be broadcast again by the computer system 300 by a remote server, and will be returned to the web browser 420 currently performed especially there so that it may be expressed with the dashed line 630. Then, the block 635 is performed, a set of a record is received, behind, URL is extracted from the record to each record, a website is accessed by the URL, and the related file which contains the document corresponding to the record further is downloaded from there. Download of all the documents will perform the block 640. This block extracts all the texts which contain all the texts in the HTML tag relevant to that document first from that document to such each document. Then, in order to make easy the natural language processing performed to one sentence at once, the text about each document is cut down by text file which occupies the separate line in a file of each text (or question) by the conventional one-sentence logging processing. Then, the block 640 receives each line of the text of the document, The NLP routine 700 (this is later explained in detail in relation to drawing 7) is repeated and called, each of these documents is analyzed, a set of the logical form triad of the correspondence about each line of the text of the document is built, and it memorizes locally. Although the operation in the block 645 was explained as the thing in the blocks 610, 635, and 640, and a thing performed in parallel fundamentally, Operation of the former block may be performed in front of them or to either the back one by one with operation of the blocks 610, 635, and 640 based on the conditions in actual mounting. Like [in the case of another example of an invention of us who replace with this and explain later in relation to drawing 13 B from drawing 10], The logical form triad about each document is calculated beforehand, and is memorized, it may be used at the time of search of a next document, and these triads are only accessed in this case, without being calculated at the time of search of a document. In this case, these triads are a certain modes and will be memorized as another entry by either of the data sets which contain the record about that document, or its document as a property (attribute) of that stored document, for example.

[0052]

Anyway, it returns to the process 600 shown in drawing 6 A and drawing 6 B, and the block 650 will be performed, if a set of a logical form triad is built to each both sides of the searched document in a set of a query and an output document and is memorized thoroughly. This block compares each of the logical form triad of a query with each of the logical form triad about each of the searched document, and traces coincidence with one either one triad of the queries or triad of the documents. The form of coincidence as an example is defined as obtaining the same coincidence between these two triads in the point of the both sides of the node word and the related child type between these triads. Especially In the case of one pair of

logical form triads of illustration, i.e., "word 1a-relation child 1-word 2i and 1d of word-relation child 2-word 2b." Only when the word 1a and the word 1b of a node word are the same, the word 2a and word 2b of a node word are the same and the related child 1 and the related child 2 are the same, coincidence takes place. If not all the three elements of one triad are identically [to the element to which another triad corresponds] in agreement, these two triads are not in agreement. If the block 650 is completed, the block 655 is performed, a triad in agreement is not obtained, i.e., all the searched documents without the triad of a query and a triad in agreement will be canceled. Then, the block 660 is performed. A score is assigned to all the remaining documents based on those dignity that exists with the block 660 about the type of the related child of a triad in agreement, and each of these documents. Although shown in Table 800 of drawing 8 A, the dignity of correspondence [like] is assigned to each of the related child of a different type which may be especially produced in a logical form triad. For example, a static number of dignity which is called 100, 75, 10, and 10 and which was defined beforehand can be assigned to the related children Dobj, Dsub, and Ops and Nadj of illustration, respectively so that it may be illustrated. When dignity shows the coincidence on the right meaning of a query and a document, it reflects the relative importance considered to attribute to the related child. Generally the actual numerical value of such dignity is defined based on experience. As it explains in detail in relation to drawing 8 B later, it is the function by which the score was beforehand defined about each of the remaining documents, and is the sum of the numerical value of the dignity of the triad (all the triads that are doubly in agreement ignore.) which is in agreement with the meaning as an example here. In this way, once weighting of the document is carried out, the block 665 is performed and a document is ranked as the descending order of a score. Finally, the block 670 is performed and 5 to ten documents are displayed on the document of the small group who shows the typical highest score and who was specified beforehand, and a type target in order of a rank. Then, when a user "clicks" a user's mouse suitably on the button of the correspondence displayed by the web browser 420, for example, The next group of the document ranked as the computer system (client PC) 300 is displayed, and like the following, this is continued until it fully examines all the documents as which the user was ranked one by one, and the process 600 is completed there.

[0053]

Drawing 7 shows the flow chart of the NLP routine 700. This routine can give the input text of one line, and also when that text is a sentence in a query and a document, or a part of any of a text, it builds the logical form triad of the correspondence about it.

[0054]

The block 710 processes the line of an input text first, and generates logical form graphs, such as the graph 515 etc. of the illustration shown in drawing 5 A, at the same time it goes into the

routine 700 especially. A logical form graph is computed from this syntax-analysis tree behind including the processing on the morpheme syntax in which this processing generates an syntax-analysis tree. Then, as shown in drawing 7, the block 720 is performed and a set of the logical form triad of correspondence is extracted from a graph (it reads). If this is performed, the block 730 will be performed and each of such a logical form triad will be generated as a separate and distinguished text character sequence by which formatting was carried out. Finally, the block 740 is performed and a set of the logical form triad about the line is memorized as a series of text character sequences by which memorized the line of the input text to the data set (namely, database), and formatting was carried out to it. If this set is memorized thoroughly, executive operation will come out of the block 700. When it replaces with this, and a different expression like a logical form graph is related with a logical form and used about our invention instead of a logical form triad, for example, The blocks 720 and 730 can be easily changed so that the specific form may be generated as a character string by which formatting was carried out, and the block 740 can be changed so that the form may be memorized instead of the logical form triad to a data set.

[0055]

In order to fully recognize the mode of the invention of us which compares coincidence of a logical form triad, makes weighting it, and ranks the document of correspondence further, drawing 8 B is referred to. This figure is shown by a diagram and comparison of the logical form triad according to instruction of an invention of us, memory of a document, a rank, and a selection process these, In the blocks 650, 660, 665, and 670 shown altogether, it is carried out to drawing 6 A and drawing 6 B about the query of illustration, and the group of illustration of three searched documents. For the purpose of illustration, it assumes that the user gave the query 810 of the full text to the search system of the invention of us, and this query presupposes that it is what is called "How many hearts doesan octopus have?." This query should be answered and the three documents 820 should be eventually searched by the statistical search engine. The 1st document (it is described as the document 1.) is a recipe containing artichokehearts and octopus among these documents. The 2nd document (it is described as the document 2.) is a paper about octopi (a general octopus). The 3rd document (it is described as the document 3.) is a paper about deer (deer). These three documents and queries are changed into the logical form triad of those components, and the processing about them is comprehensively expressed by "NLP" (natural language processing). The logical form triad about a query, the document 1, the document 2, and the document 3 obtained as a result is given to the blocks 830, 840, 850, and 860, respectively.

[0056]

Once these triads are defined in this way, so that it may be expressed with the dashed lines 845, 855, and 865, The logical form triad about a query is compared with the logical form triad

about the document 1, the document 2, and the document 3 one by one, respectively, and it is checked whether one of documents contains one logical form triad of the queries and a triad in agreement. Like [in the case of the document 1], the document which does not contain such a triad in agreement is canceled, and is not further taken into consideration. On the other hand, the document 2 and the document 3 contain a triad in agreement. Especially the document 2 contains such three triads. That is, these are "HAVE-Dsub-OCTOPUS" relevant to a sentence different from "HAVE-Dsub-OCTOPUS" and "HAVE-Dsub-HEART" relevant to one sentence, for example (the sentence of these is not shown specifically). Two of these triads are the same, namely, it is "HAVE-Dsub-OCTOPUS." The score about a document is the sum of the numerical value of the dignity of all the triads that are in agreement with a meaning in the document, for example. All the triads that are doubly in agreement are disregarded about all the documents. The illustration rank of the relative dignity of the related child of a different type which may be produced in a triad, To the order of dignity to the biggest, small dignity, the beginnings are the combination (Dobj) of a verb-object, the combination (Dsub) of a verb-subject, a preposition, and a functional word (for example, it is a modifier (for example, Nadj) at (Ops) and the last.). Such a weighting method is shown in the triad weighting table 800 of the illustration shown in drawing 8 A. In order to simplify this figure, Table 800 shows only the thing relevant to the triad shown in drawing 8 B excluding all various related children that may arise in a logical form triad. These metrics have given the check ("RE") mark to the specific triad which contributes to that score among each document. Of course, the thing done [not adding dignity, in order to use for a document except what we chose as metrics for carrying out score attachment specified beforehand, for example, to improve the selectivity (distinction) of a document] for multiplication, Or it may be removing the dignity of other triads other than adding dignity by another ******, for example, include two or more coincidence of the same type, and/or **** etc. The score will take the following into consideration in a certain mode about arbitrary documents. Namely, frequency or semantic contents of the node words of the triad in the document itself, and these node words in the document, They are the frequency as the specific logical form (or paraphrase of that) in the frequency of the word of the specific node in the document, semantic contents, or its document, and/or the specific whole logical form triad, and the length of the document.

[0057]

Therefore, if the dignity specified in the metrics of score attachment and Table 800 of drawing 8 A which we chose is taken into consideration, The score of the document 2 is 175 and this is formed by combining the dignity, i.e., 100 and 75, of the first two triads relevant to the 1st sentence shown in the block 850 in a document. In relation to that 2nd sentence in this document, it is the 3rd triad indicated to this block, and what is already in agreement with one of the triads of other which exist in a document is disregarded. Similarly, the score about the

document 3 is 100 and this is formed of 100 at the dignity about the triad which corresponds within this specific document as indicated to the block 860, i.e., here. Based on a score, the document 2 is ranked before the document 3 and a user is shown these documents in the turn. although it did not happen here, when any two documents have the same score, these documents are ranked and carried out in the same turn as being provided by the conventional statistical search engine, and a user is shown them in the turn.

[0058]

It will be easily understood that distributed **** is also good for various computers which form an information retrieval system as a whole, even if various portions of the processing used in order to realize our this invention exist in a single computer, if it is a person skilled in the art clearly. Drawing 9 A to drawing 9 C shows three examples from which the information retrieval system which adopted instruction of our this invention differed about this point, respectively.

[0059]

One of such the alternative examples is shown in drawing 9 A, and all the processings are performed by the single local computers 910, such as PC, here. In this case, the computer 910 acts as a host of the search engine, and with that engine. The query of the full text to which the input document was given by the indexing and the user is answered (whether it is locally placed there by CD-ROM or other storages.). Or the set of the searched document which is accessible to the computer, searchs a certain data set, and forms an output document group is generated eventually. This computer acts as a host of the processing of an invention of us further, and both a query and such each document are analyzed, A set of the logical form triad of correspondence is generated, a set of a triad is compared after that, in an above-mentioned mode, a document is score attachment ** ranked, and, finally it is arranged there, for example, or a result is shown to a local user accessible there.

[0060]

Another alternative example is shown in drawing 9 B, this drawing 9 B includes the specific situation shown in drawing 2, and a search system is formed here with client PC by which network connection was carried out to the remote server. Here, client PC920 is connected to the remote computer (server) 930 by the network connection 925. The user who is in client PC920 inputs the query of the full text, and PC transmits this to a remote server via network connection. Client PC analyzes a query further and generates a set of the logical form triad of the correspondence. A server acts, for example as a host of the conventional statistical search engine, as a result, answers this query, performs statistical search, and generates a set of a document record. And a server returns a set of a record and returns each document in a set of an output document to client PC autonomously eventually based on the capability of the command of a client, a search engine, or related software. And client PC analyzes each of the document of the correspondence received in the set of an output document, and generates a

set of the logical form triad about it.Client PC ranks second, compares a set of a triad appropriately, in an above-mentioned mode, chooses and carries out score attachment of the document, ranks it, finally shows a local user a result, and completes the processing.

[0061]

The further example is shown in drawing 9 C. Although this example uses the same physical hardware and network connection as drawing 9 B, Client PC920 accepts the request of the full text query from a local user, and transmits the request of the query to the remote computer (server) 930 via the network connection 925. This server provides the natural language processing according to this invention rather than only acts as a host of the conventional search engine. In this case, not client PC but a server will analyze a query appropriately, and will produce the set of correspondence of the logical form triad for it. If the server is required again, it will download each searched document in an output document group, then it will analyze such each document, and will generate the set of correspondence of the logical form triad for it. Then, a server will compare a set and document of the triad for a query appropriately, will choose a document as mentioned above, will attach a score to it, and will attach the rank. The server 930 will transmit the remaining search documents to client PC920 via the network connection 925 in order of a rank, and will be made to display them there, once this rank is performed. A server transmits these documents for every group according to a user's directions as mentioned above, or in order to choose them for every group and to make it display with client PC, it can transmit all the documents one by one.

[0062]

The remote computer (server) 930 may be a distributed processing method which does not need to be realized by one computer which gives the above-mentioned conventional retrieval processings, natural language processing, and all the processings of relation, and is shown in drawing 9 D. In that case, the processing which this server contracts is distributed between the individual servers in a distributed processing method. Here, the server 930 consists of the front-end processor 940 which distributes a message via the connection 950 to a series of (the server 1, the server 2, --, the server n are included) servers 960. Each of these servers carries out the specific portion of the process of this invention. At this point, since the indexing of an input document is performed for next search and it stores in the data set on a large capacity data storage, the server 1 can be used. The server 2 can realize a search engine like the conventional statistical engine for answering the query given by the user seen off by the front-end processor 940, and pulling out the document record of a lot from a large capacity data storage. These records will be sent to the server n via the front-end processor 940 from the server 2, and post-processing of downloading each document of the correspondence from the website or database of correspondence during an output document group will be performed. The front-end processor 940 will send the query to the server n again. The server n analyzes

the query and each document appropriately next, and produces a set of correspondence of a logical form triad, Next, a set of a triad is compared appropriately, a document is chosen as mentioned above, a score is attached to it, a rank is attached, the document ranked after that is returned to client PC920 via the front-end processor 940, and a rank is made to be indicated there. of course, various operations used in processing of this invention be static according to the conditions produced by the conditions produced at the time of execution, and/or others -- be dynamic -- it may distribute in the server 960 by the arbitrary methods of many other methods. Both the database for the search engine of the former [server / 930] for example, and the dictionary used for natural language processing are memorized, It is a shared direct access storage device (DASD) accessible from all the processors in a server, for example, well-known SHISUPU REXX composition (or other same distributed multi-process environment) can also realize.

[0063]

Although explained as what answers each searched document record in this invention, downloads a document, then analyzes the record locally for example, with client PC, and produces the logical form triad of the correspondence, It changes to this, and these triads may be generated while the search engine is carrying out the indexing to the document. A search engine at this point for example, when each new document for performing an indexing is found using a web crawler, An engine can download the perfect file for the document, and the document can be pretreated immediately after [the] or by analyzing the document by batch processing later further, and generating the logical form triad. The search engine will memorize these triads in the database next as some records in which the indexing of [for the document] was carried out at the time of the end of pretreatment. Behind, whenever the document record answers for example, a search query and is searched, the triad for it is returned to client PC as some document records for the purposes, such as comparison. It is effective in the processing time of most quantity in client PC being saved by pretreatment of the document within a search engine, and the throughput of a client can be increased by it.

[0064]

Although concrete use with the search engine of the Internet base was explained as an example, this invention, whether or not this invention will be (a) Internet base -- a network facility for exclusive use etc. -- accessible arbitrary networks -- an accessible search engine. (b) The local search engine which operates to the data set which itself possesses, and which was recorded beforehand, for example, the data retrieval application of the CD-ROM base represented by an encyclopedia, a yearbook, or other standalone version stand-alone data sets and/or (c) -- it is equally [to use in the arbitrary combination] applicable.

[0065]

Bearing the above in mind drawing 10 A and drawing 10 B, The example of further others of

this invention is shown collectively, and a logical form triad is generated by pretreatment of a document in this example, Make into a standalone version stand-alone data set the triad, document record, and the document itself which are produced as a result, and The existing storage, for example, one or more CD-ROMs -- or (represented by a removable hard disk, a tape, optical magnetism, mass magnetism, or the electronic storage) -- others -- the distribution to an end user becomes easy by saving collectively to the mass medium of portability. Right arrangement of these drawings is as being shown in drawing 10. By summarizing the retrieval application itself and the data set which accompany it and which should be searched to a common medium, and putting it in, In order to obtain stand-alone data retrieval application and to search a document by it, it is less necessary to carry out network connection to a remote server.

[0066]

As illustrated, this example consists of three portions intrinsically [document indexing partial 1005_1, duplicate partial 1005_2, and user partial 1005_3]. Partial 1005_1 collects and carries out the indexing of the document, creates the data set 1030, i.e., the data set to illustrate, and the data set 1030, The document repository for the collection of standalone version document retrieval application, for example, an encyclopedia, a yearbook, a private library (it is (like law reports)), and periodicals, etc. is formed. The cost for reproducing the medium of CD-ROM which has mass storage capacity, and other gestalten is falling quickly, and this example is attractive especially in order to distribute a lot of documents with sufficient cost-performance to a large user community with the performance which searchs it correctly.

[0067]

Anyway, the documents inputted in order to carry out an indexing and to form a data set are collected from sauce with various any number, and are given to the computer 1010 one by one. With the suitable software memorized in the memory 1015, realize this computer and a document indexing engine a document indexing engine, The suitable entry which creates the record for such each document in the data set 1030, and saves information on the record for the document, and includes the copy of the document itself is created and saved in a data set. The engine 1015 performs the triad generation process 1100. This process explained in detail below in relation to drawing 11 is separately performed for each [an indexing is carried out] document of every. Intrinsically this process with having mentioned above to the block 640 shown in drawing 6 A and drawing 6 B in a similar manner intrinsically, By analyzing and doing the text phrase in a document so, a set of correspondence of the logical form triad to the document is constituted, and it memorizes in the data set 1030. Since each of all the processes of the others which are performed with the indexing engine 1010 shown in drawing 10 A and drawing 10 B, and attach an index to a document, for example, processes of generating the suitable record for it, is unrelated to this invention, they are not described in

detail. Once a set of a triad is generated by the process 1100, it is enough just to say that the engine 1015 memorizes this set to the data set 1030 with the copy of the document itself and the document record made to it. Therefore, after all the indexing operations finish, the data set 1030 not only has memorized in it the perfect copy by which the indexing was carried out and that of a sake of all the documents, but has memorized the set of the logical form triad for the document.

[0068]

Once the indexing of all the desired documents is carried out appropriately, although this can regard the data set 1030 as a "master-data set", it will be reproduced by duplicate partial 1005 $_2$ next. Within partial 1005 $_2$, the conventional medium duplication system 1040 the copy of the contents of the master-data set supplied by the line 1035, With the copy of the suitable file for the retrieval software containing the retrieval process and user install program which are supplied by the line 1043. It writes in a common storage like one or more CD-ROMs repeatedly, and stand-alone document retrieval application is formed collectively. Each duplicate 1050 $_1$, 1050 $_2$, --, a series of medium duplicates 1050 that have 1050 n are generated by the system 1040. All the duplicates include the copy of the document search application file supplied by the line 1043, and the copy of the data set 1030 supplied by the line 1035, as it is the same and duplicate 1050 $_1$ is shown concretely. According to the size of a data set, and composition, each duplicate may straddle one or more separate media, for example, separate CD-ROM. Behind, typically, a duplicate circulates all over a user community, as the dashed line 1055 shows by acquisition of a license.

[0069]

As a user, for example, user $_j$, is shown in user partial 1005 $_3$ (CD-ROM1060 is shown), once it obtains a duplicate like CD-ROM$_j$, A user according to the computer system (it is (like [though it is not the same composition] PC which has composition like client PC300 substantially shown in drawing 3)) 1070. Document retrieval application containing this invention can be performed to the data set memorized by CD-ROM$_j$, and a desired document can be pulled out from there. After CD-ROM$_j$ comes to hand, insert [especially a user / CD-ROM / in PC1070], he begins to execute the install program memorized by CD-ROM, and by it. The copy of a document search application file is made, it is installed in the memory 1075 of PC, and the directory in a hard disk usually specified beforehand, and the document retrieval application 1085 is created on PC by it. This application includes the search engine 1090 and the retrieval process 1200. If installation is once completed and the application 1085 is called, the user can search the data set of CD-ROM$_j$ by giving the query of the suitable full text to application. Answering a query, a search engine pulls out a set of the document containing the logical form

triad memorized for that for those documents, and such each document from a data set. A query is also given to the retrieval process 1200. This process is dramatically similar to the retrieval process 600 mentioned above in relation to drawing 6 A and drawing 6 B, and analyzes a query, therefore constitutes a logical form triad. Then, the process 1200 shown in drawing 10 A and drawing 10 B compares the logical form triad for [each] the searched document in the set, especially the record for it with the triad for a query. Based on coincidence of the triad generated among them, and those dignity, the process 1200 carries out score attachment of each of the document which shows at least one triad in agreement in the mode mentioned above in detail, These documents are ranked by the score of a descending order, and the document record of small groups typically fewer than 5-20 or it who finally have the highest rank is visually shown to a user. The user can examine that of these and can direct to search and display the whole copy of the arbitrary documents considered that there is relation to document retrieval application. Once a user examines the first group's document record to the first group's search document, The user can demand the document record of the next group who has a rank high next, and he can perform this until he finishes examining all the searched document records like the following. Although the application 1085 answers a query and returns the ranked document record in an initial state, it changes to this, and this application may answer a query and may return the copy as which the document itself was ranked.

[0070]

Drawing 11 shows the triad generation process 1100 performed with the document indexing engine 1015 shown in drawing 10 A and drawing 10 B. As mentioned above, pretreatment of the document to which the indexing of the process 1100 should be carried out, It carries out by analyzing and doing so the text phrase in the document by constituting a set of correspondence of the logical form triad for the document, and memorizing in the data set 1030. The block 1110 will be performed if the process 1100 is started especially. This block extracts all the texts containing the arbitrary texts which are in the HTML tag relevant to that document first from that document. Then, in order to make easy the natural language processing performed one sentence at a time at once, the text for each document is disassembled by the conventional one-sentence logging processing, and each sentence (or interrogative sentence) serves as a text file which occupies a separate line within a file. Then, the block 1110 calls separately the NLP (in relation to drawing 13 A, it mentions later in detail) routine 1300 for every line of the text in the document, This document is analyzed, a set of correspondence of the logical form triad for that line is constituted, and it memorizes locally in the data set 1030. If these operations are completed, execution of the block 1110 and the process 1100 will be completed.

[0071]

A flow chart of the retrieval process 1200 of this invention which is used in the concrete example of this invention shown in drawing 10 A and drawing 10 B is collectively shown in drawing 12 A and drawing 12 B. The right arrangement of the drawing of drawing 12 A and drawing 12 B is as being shown in drawing 12. (It was shown in drawing 6 A and drawing 6 B, and mentioned above in detail) In PC1070 (refer to drawing 10 A and drawing 10 B), all the operations shown in drawing 12 A and drawing 12 B are performed by contrast [the retrieval process 600] a common computer system and here. In order to help an understanding, please refer to simultaneously drawing 10 A and drawing 10 B in the following explanation.

[0072]

A start of the process 1200 will perform the block 1205 first. When this block is performed, it makes a user input the query of the full text. Once this query is obtained, an execution path will branch and will progress to the course 1245 according to the block 1210 and the course 1243 by the course 1207. The block 1245 calls the NLP routine 1350 as performing, analyzes a query, constitutes a set of the logical form triad of correspondence, and memorizes it in the memory 1075 locally. If the block 1210 is performed, as the dashed line 1215 shows, it will transmit the query of the full text to the search engine 1090. A search engine answers a query, performs the block 1220, and searches both a set of a document record, and the logical form triad of the relation relevant to each of such a record at this time. If this set and the logical form triad of relation are searched, both will be given to the process 1200 as the dashed line 1230 shows, and will specifically be given to the block 1240 in there. The block 1240 only receives this information from the search engine 1090, and in order to use it behind, it memorizes it in the memory 1075. Although the operations in the block 1245 were explained to be the operation in the blocks 1210, 1090, and 1220, and a thing intrinsically performed in parallel, Operation in the block 1245 may be performed from a viewpoint on actual execution in in-series before the operation in the block 1210, 1090, or 1220, or to the back.

[0073]

The block 1250 will be performed if a set of a logical form triad is memorized to the memory 1075 for both a query and each searched document record. This block traces coincidence between the arbitrary triads in a query, and the arbitrary triads of the arbitrary documents of correspondence as compared with each of the logical form triad for each document record which was mentioned above in detail and which is a mode and was searched in each of the logical form triad in a query. That is [once the block 1250 is completed, / the block 1255 is performed and a triad in agreement is not shown], all the searched records to the document which does not have the arbitrary triads in a query and a triad in agreement are discarded. Then, the block 1260 is performed. All the document records which remain with the block 1260 as mentioned above, If a score can be assigned based on the types and those dignity of a relation of the triad in agreement which exists for every document of correspondence and

weighting of the document record is carried out such, the block 1265 is performed and a record is ranked as the descending order of a score. Finally, the block 1270 is performed and a record is displayed on the small group and type target which show the typical highest score and which were specified beforehand in order of a rank about the document record of 5 or 10. Then, a user by for example, the thing a mouse "is clicked" appropriately on the button of the correspondence currently shown by the computer system 1070. The next group of the ranked document record is displayed on the system, and the operation is performed until it fully investigates in order all the document records in which the user was ranked like the following (and the interested arbitrary documents in it are accessed and it is investigated). At this time, it is completed and execution ends the process 1200.

[0074]

Drawing 13 A shows the flow chart of the NLP routine 1300 performed within the triad generation process 1100 shown in drawing 11. As mentioned above, the document to which the indexing of the NLP routine 1300 should be carried out and which carries out ingress, One line of a text is specifically therefore analyzed, a set of correspondence of the logical form triad for the document is constituted, and it is memorized in the data set locally shown in drawing 10 A and drawing 10 B. The routine 1300 is shown in drawing 7 and operates in a similar manner intrinsically with the NLP routine 700 mentioned above in detail.

[0075]

If the routine 1300 is started especially, the block 1310 will be performed first and will generate a logical form graph like the graph 515 of the illustration which processes the line of an input text and is shown in drawing 5 A. Then, as shown in drawing 13 A, the block 1320 is performed and a set of the logical form triad of correspondence is extracted from the graph (it reads).Once this happens, the block 1330 will be performed and each of such a logical form triad will be separately generated as a text character sequence by which formatting was distinguished and carried out. Finally, the block 1340 is performed and a set of the logical form triad for the line is saved at the data set 1030 as the line of the inputted text, and a series of text character sequences by which formatting was carried out. If this set is memorized thoroughly, execution of the block 1300 will be ended. If it changes to this and not a logical form triad but a different form, for example, a logical form graph, or its subgraph is used in relation to this invention, It could change easily instead of the block 1340 being a logical form triad so that the form may be memorized to a data set, so that the specific form may be generated as a character string by which formatting was carried out in the blocks 1320 and 1330.

[0076]

Drawing 13 B shows the flow chart of the NLP routine 1350 performed within the retrieval process 1200. As mentioned above, by user $_j$, the NLP routine 1350 analyzes the query given

to the document (shown in drawing 10 A and drawing 10 B) retrieval application 1085, constitutes a set of the logical form triad of correspondence for it, and memorizes it locally in the memory 1075. The only difference on operation between the routine 1300 and the routine 1350 which were mentioned above in detail in relation to drawing 13 A is a place where the triad of correspondence is memorized. That is, it is the point that the data set 1030 memorizes in execution of the block 1340 in the NLP routine 1300, and the memory 1075 memorizes in execution of the block 1390 in the NLP routine 1350. With other blocks of the routine 1350 and operation specifically performed by the blocks 1360, 1370, and 1380 being performed by the blocks 1310, 1320, and 1330 of the routine 1300, respectively, since it is substantially the same, Detailed explanation of the former block is ******(ed).

[0077]

In order to try the performance of the retrieval process of this invention which was generally mentioned above in relation to drawing 1 A in a tentative way, in the search system of this invention, the ALTA VISTA search engine was used as a search engine. On the Internet, this engine accessible in everyone is the conventional statistical search engine called that the indexing of the web page exceeding 31 million is carried out, and is used widely (28 million hits are recorded per day at estimate now.). the retrieval process 600 of this invention -- MICROSOFT OFFICE -- it is contained in the syntax checker who accomplishes a sweet part 97 **********. It realized on PC of general Pentium 90 MHz using various natural-language-processing components containing a dictionary file ("OFFICE" and "OFFICE 97" are the trademarks of Microsoft Corporation of Redmond, Washington). We used the on-line pipeline processing model. That is, while the user was waiting for the continuing result, the document was collected and processed on-line in the pipeline's mode. This specific PC took about 1/2 second to generate a logical form triad for every sentence from about 1/3 second.

[0078]

The volunteer was requested to make the query of the full text for giving a search engine. A total of 121 wide range, mutually different queries are made, The typical thing "Why was the Celtic civilization so easily conquered by the Romans? (why was Celt civilization conquered by the Roman simply?)", "Why do antibiotics work on colds but noton viruses? (isn't it effective against a virus that an antibiotic is [why] effective against cold?)", "Who isthe governor of Washington? (the governor of Washington is someone ?)", "Where does the Nile cross the equator? (it is somewhere ? that the Nile River intersects the equator)", And it was called "When didthey start vaccinating for small pox? (about what time is it that the vaccination of variola was able to begin?)." Each of these 121 queries were given to the ALTA VISTA search engine, and top 30 document which consists of an available thing which answered each query and returned was obtained. Only less than 30 document returns from some queries, and all the documents which returned in that case were used. A total of 3361 documents (namely, "raw"

document) were obtained to all the 121 queries.

[0079]

Each of the document of 3361 and the query of 121 was analyzed by the process of this invention, and the set of correspondence of a logical form triad was generated. It was compared appropriately, the document of the result was chosen as mentioned above, and the set had the score and the rank attached.

[0080]

All the documents of 3361 were manually evaluated separately about relevance with the query of the correspondence from which the document was obtained. In order to evaluate relevance, the person of one person who does not know the concrete purpose of experimenting in an artificer was used as an evaluator, and about relation with the query of the correspondence, each of these 3361 documents was manually ranked subjectively as being "optimum", "it being related", or "with no relation", and was carried out. It was presupposed that the optimal document was a thing including the clear reply to the query of correspondence. It was presupposed that it was a certain thing of relevance although the document with relation does not include the clear reply to a query. That whose unrelated document is not the useful reply to a query, That is, it was presupposed from URL of the correspondence which there is no relation in a query, was based on languages other than English, or was given with the ALTA VISTA engine (namely, "cobweb" link) that it was a document which cannot be searched. In order to raise evaluation accuracy, the second evaluator on the subset in these 3361 documents, and a concrete target. The document (431 among 3361 documents) which showed at least one logical form triad which is in agreement with the logical form triad in the query of the correspondence, Although it was ranked that it is related till then or the optimal, the document (102 among 3361 documents) which does not have a logical form triad in agreement was investigated. It was examined by the third evaluator who turns into an "arbitrator" when there was a difference of opinion of these ranks to a document.

[0081]

As a result of this experiment, it was observed in all the related documents that the search system of this invention showed the improvement rather than the raw document which an ALTA VISTA search engine returns. In the whole accuracy (namely, all the selected documents), it has improved about 200% from about 16% to about 47%, and has improved about 100% from about 26% to about 51% within top 5 documents. In addition, the accuracy of the first document returned for being the optimal has improved about 113% from about 17% to about 35% to it of a raw document by use of the system of this invention.

[0082]

In this invention, although use with a statistical search engine was concretely explained as an example, this invention is not limited to it. In the point, this invention can be used so that the

search document substantially obtained by what type of search engine may also be processed and the accuracy of the engine may be raised.

[0083]

Instead of using the dignity fixed for every attributes of various kinds of in a logical form triad, such dignity is changed dynamically and, as a matter of fact, it is good also as accommodative. In order to attain this, a study mechanism like Bayes's network or a neural network may be incorporated suitable for the process of this invention, for example, and the weight numerical value for various kinds of logical form triads may be changed into the optimal value based on study experience.

[0084]

Although it needed for the logical form triad of the process of this invention to correspond correctly, the standard of coincidence judgment is eased and it may be made to consider that a paraphrase is also coincidence for the purpose of identifying a semantic content fully similar between triads. A paraphrase may be lexical and may be structural. Probably, the example of a lexical paraphrase is a superordinate word or a synonym. The example of a structural paraphrase is use of a noun or a relative clause in coordinate relation. For example, it should be considered that the noun composition in coordinate relation like "the President and Bill Clinton" is what is done as relative clause composition like "Bill Clinton who is the President" one. About how two words are semantically similar mutual, in a semantic level, can make a fine judgment and by it. Coincidence between corpus (example) sentences which are referred to as "A tropical mountainous district may be sufficient as coffee, and it is grown" with the query "where coffee is grown" can be accepted. In addition, the procedure for judging whether coincidence exists or not can be changed with the type of the query which was able to be given. For example, it will be required that the "place" attribute must exist in the arbitrary triads relevant to the sentence currently tested, in order to consider that a certain sentence of the procedure corresponds with a query, if the query is asking about the place where something exists. Therefore, it is comprehensively specified that it includes what only the same coincidence is not only called "coincidence" of a logical form triad, but is produced from all the coincidence conditions that include such eased judgment, and that were changed.

[0085]

This invention can be easily combined with other treatment technique centering on search of graphics, a table, an image, or non text information like others, and the whole accuracy can be raised. Generally, linguistic (based on text) depiction in the document like the sign of a figure or short explanation accompanies the contents of a non text in a document well. Therefore, it can use in order to analyze and process the process of this invention, and the linguistic depiction which often accompanies the contents of a non text in use of the natural language ingredient especially. By looking for a set of the document which shows the linguistic contents

which related to the query semantically first, next processing a set of this document about those contents of a non text, The document which has the related contents of a text and the contents of a non text using the natural-language-processing art of this invention can be searched. It changes to this, and document search may be first performed about the contents of a non text, a set of a document may be taken out, and the document which has relation by next processing a set of the document about those linguistic contents by the art of this invention may be searched.

[0086]

Although various examples which adopted instruction of this invention were illustrated and being explained in detail, if it is a person skilled in the art, these instruction will be able to be easily thought out in many of other examples used in addition.

[Brief Description of the Drawings]

[Drawing 1]It is a block diagram of the very high level of the information retrieval system 5 according to our this invention.

[Drawing 2]It is a figure showing the example using instruction of our this invention of the high level of the information retrieval system 200 of the type shown in drawing 1.

[Drawing 3]It is a block diagram which is contained in the system 200 shown in drawing 2 and in which showing the computer system 300 which is a client personal computer specifically.

[Drawing 4]It is a block diagram of the very high level in which the application program 400 executed within the computer 300 shown in drawing 3 is shown.

[Drawing 5]A to D is a figure showing the logical form element of the correspondence about various examples of correspondence of the English sentence which has various complexity, and them.

[Drawing 6]Drawing 6 is a figure showing right arrangement of the drawing of drawing 6 A and drawing 6 B, and drawing 6 A and drawing 6 B are the figures doubling and showing the flow chart of the retrieval process 600 of an invention of us.

[Drawing 7]It is a figure showing the flow chart of the NLP routine 700 performed within the process 600.

[Drawing 8 A] It is a figure showing the weighting table 800 of a logical form triad in agreement as an example.

[Drawing 8 B] . Are related with an illustration question and the group of three documents of an example searched statistically. It is a figure showing visually comparison of the logical form triad according to instruction of the invention of us performed to drawing 6 A and drawing 6 B within the blocks 650, 660, and 665 shown altogether and 670, score attachment of a document, a rank, and a selection process.

[Drawing 9]A to C is a figure showing three different examples of the information retrieval system which adopted instruction of our this invention, respectively.

D is a figure which is used in realizing another different example of our this invention and in which showing the alternative example of the remote computer (server) 930 shown in drawing 9 C.

[Drawing 10]Drawing 10 is a figure showing right arrangement of the drawing of drawing 10 A and drawing 10 B, Drawing 10 A and drawing 10 B are another examples of our this invention, and are a figure doubling and showing what the logical form triad about each document is calculated beforehand, is memorized with the document record about them, and accessed at the time of next document retrieving operation.
[Drawing 11]It is a figure showing the triad generation processing 1100 performed with the document indexing engine 1015 shown in drawing 10 A and drawing 10 B.
[Drawing 12]Drawing 12 is a figure showing right arrangement of the drawing of drawing 12 A and drawing 12 B, and drawing 12 A and drawing 12 B are the figures doubling and showing the flow chart of the retrieval processing 1200 of the invention of us performed within the computer system 300 shown in drawing 10 A and drawing 10 B.
[Drawing 13 A] It is a figure showing the flow chart of the NLP routine 1300 performed within the triad generation processing 1100.
[Drawing 13 B] It is a figure showing the flow chart of the NLP routine 1350 performed within the retrieval processing 1200.

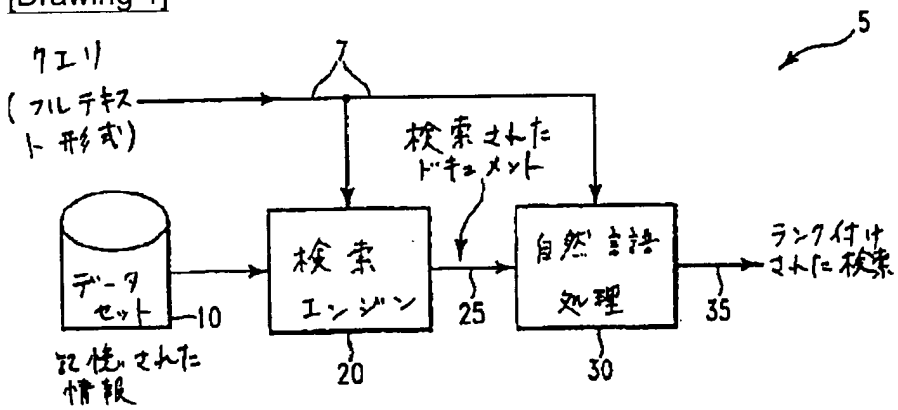
[Translation done.]

* NOTICES *

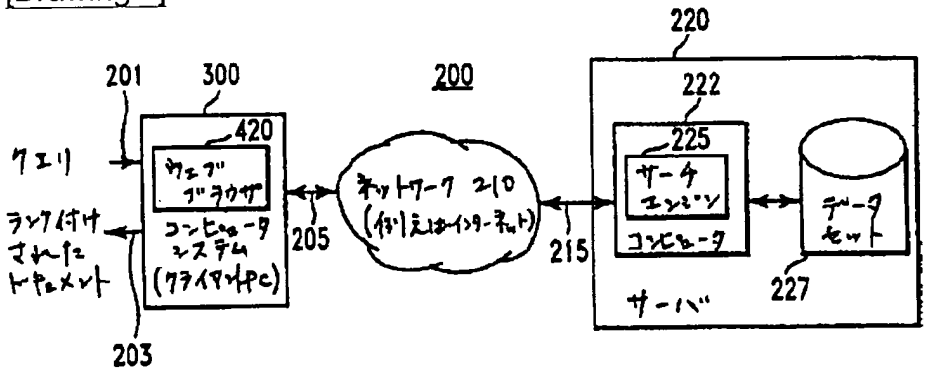
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

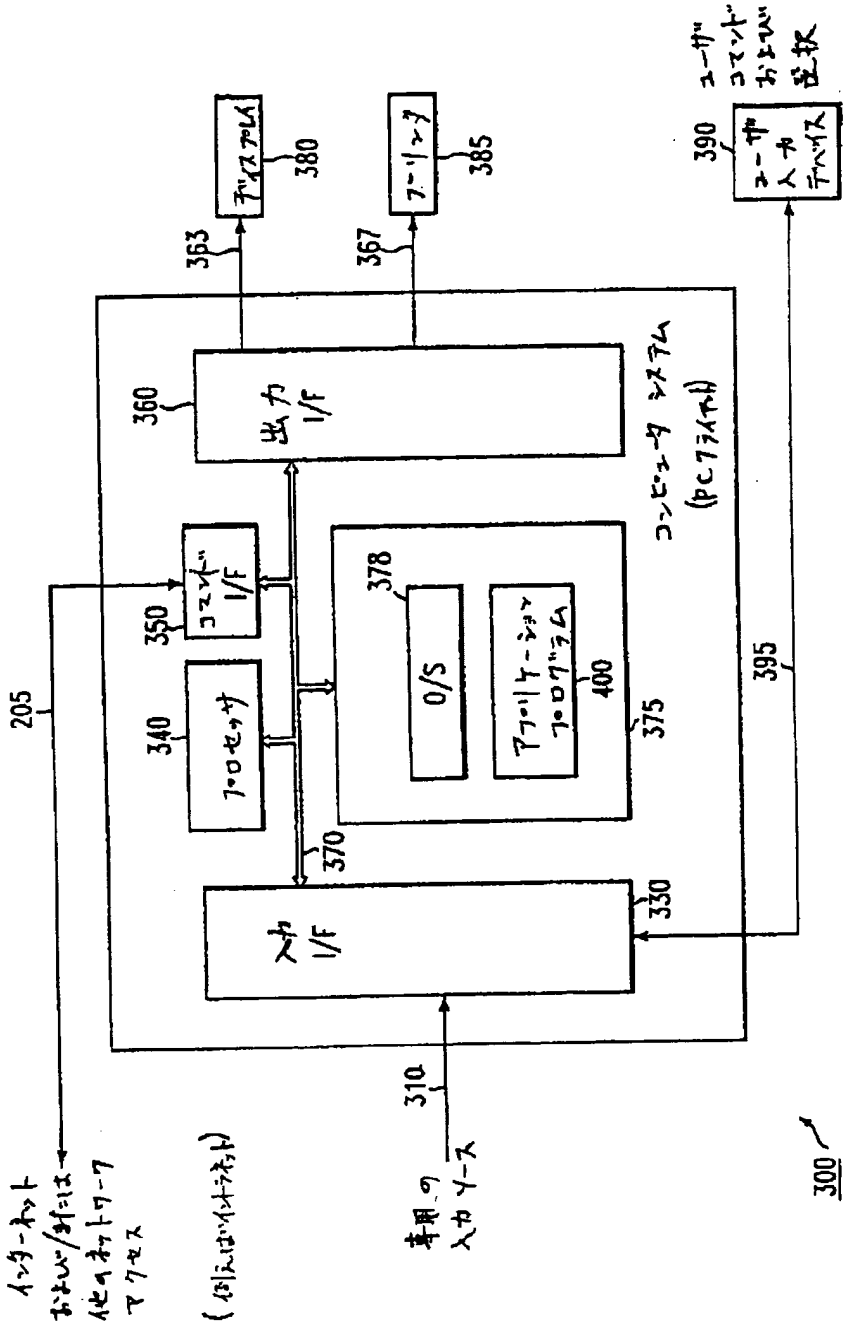
[Drawing 1]



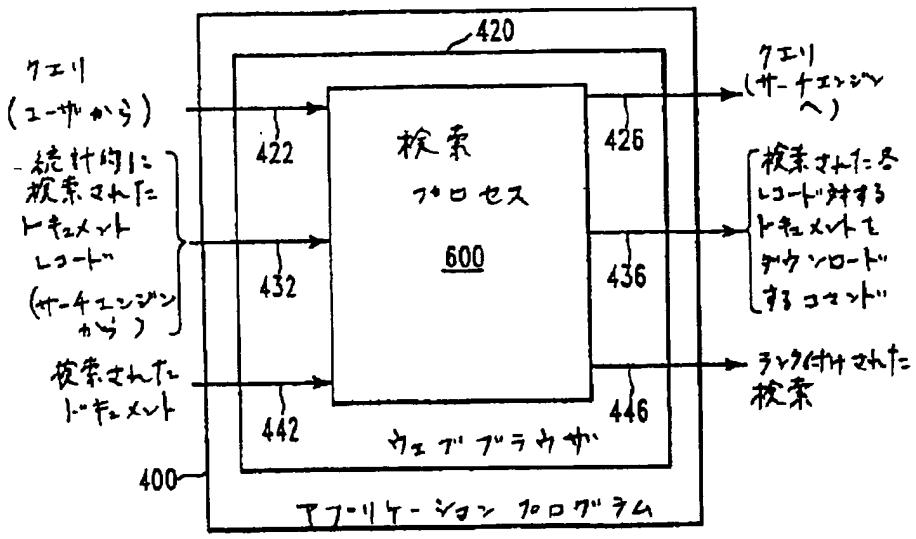
[Drawing 2]



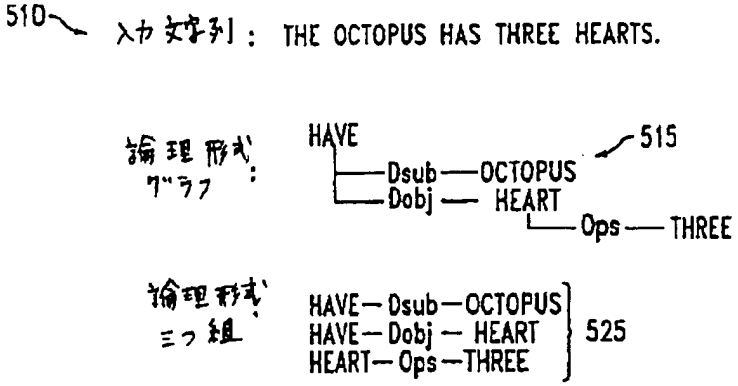
[Drawing 3]


380
385
390
363
367
360
出力 I/F
コンピュータ システム (PCプラットフォーム)
350
コマンド I/F
378
O/S
340
プロセッサ
アプリケーション プログラム
400
375
370
395
205
入力 I/F
330
310
300

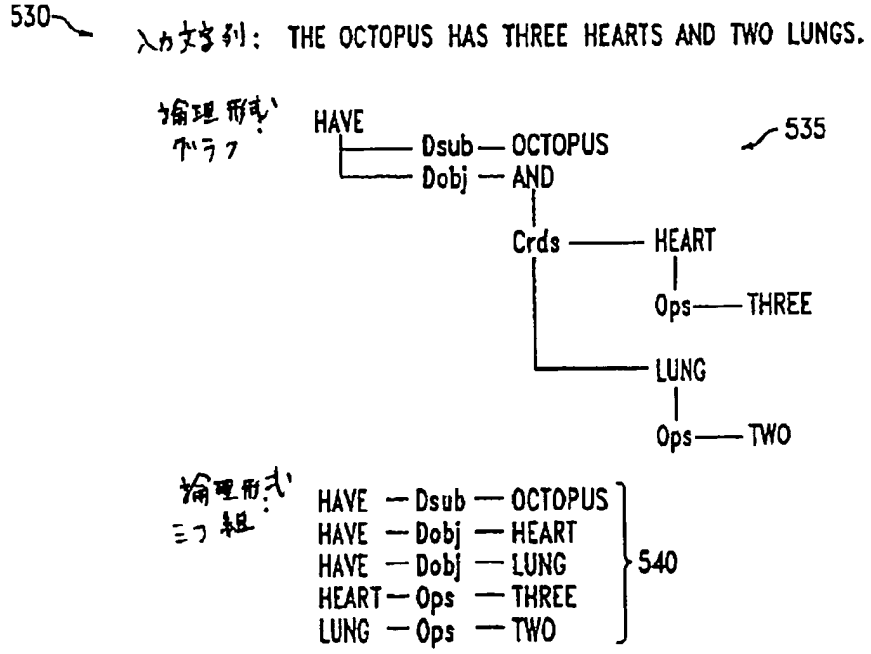
[Drawing 4]



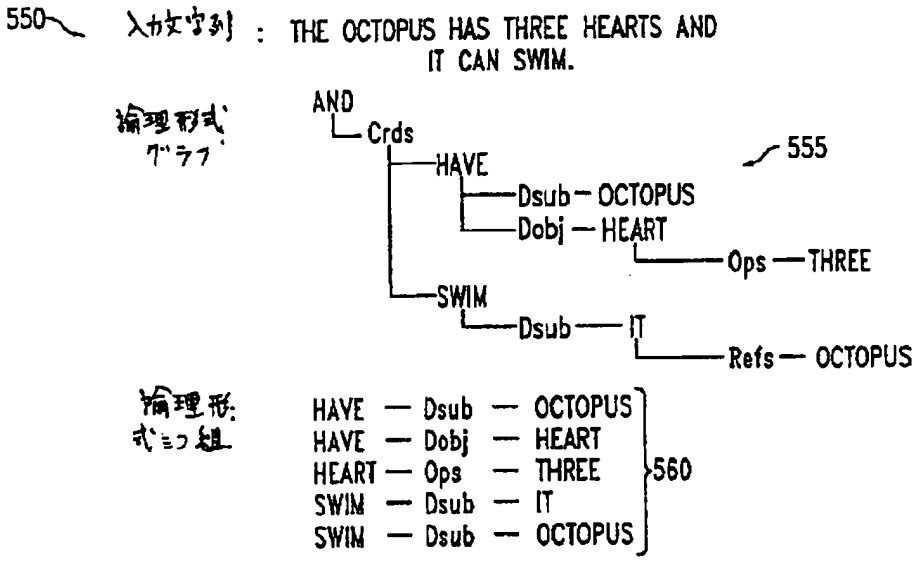
[Drawing 5 A]



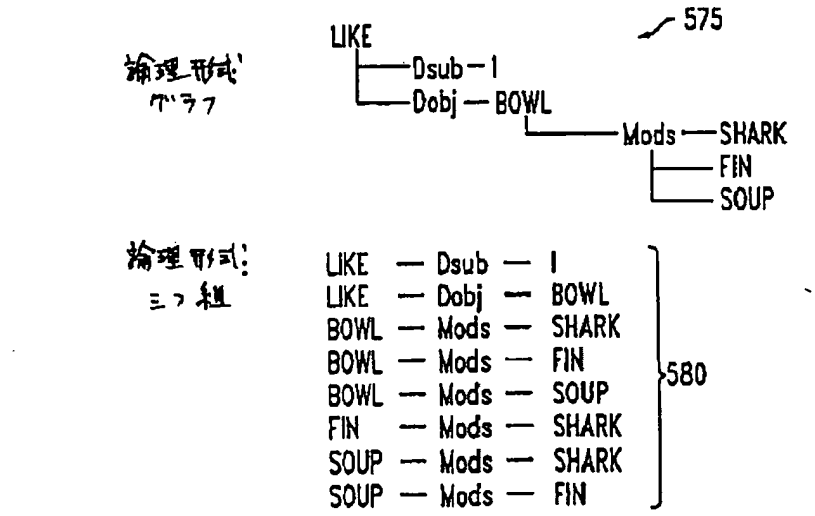
[Drawing 5 B]



[Drawing 5 C]



[Drawing 5 D]



[Drawing 6]

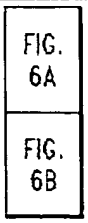
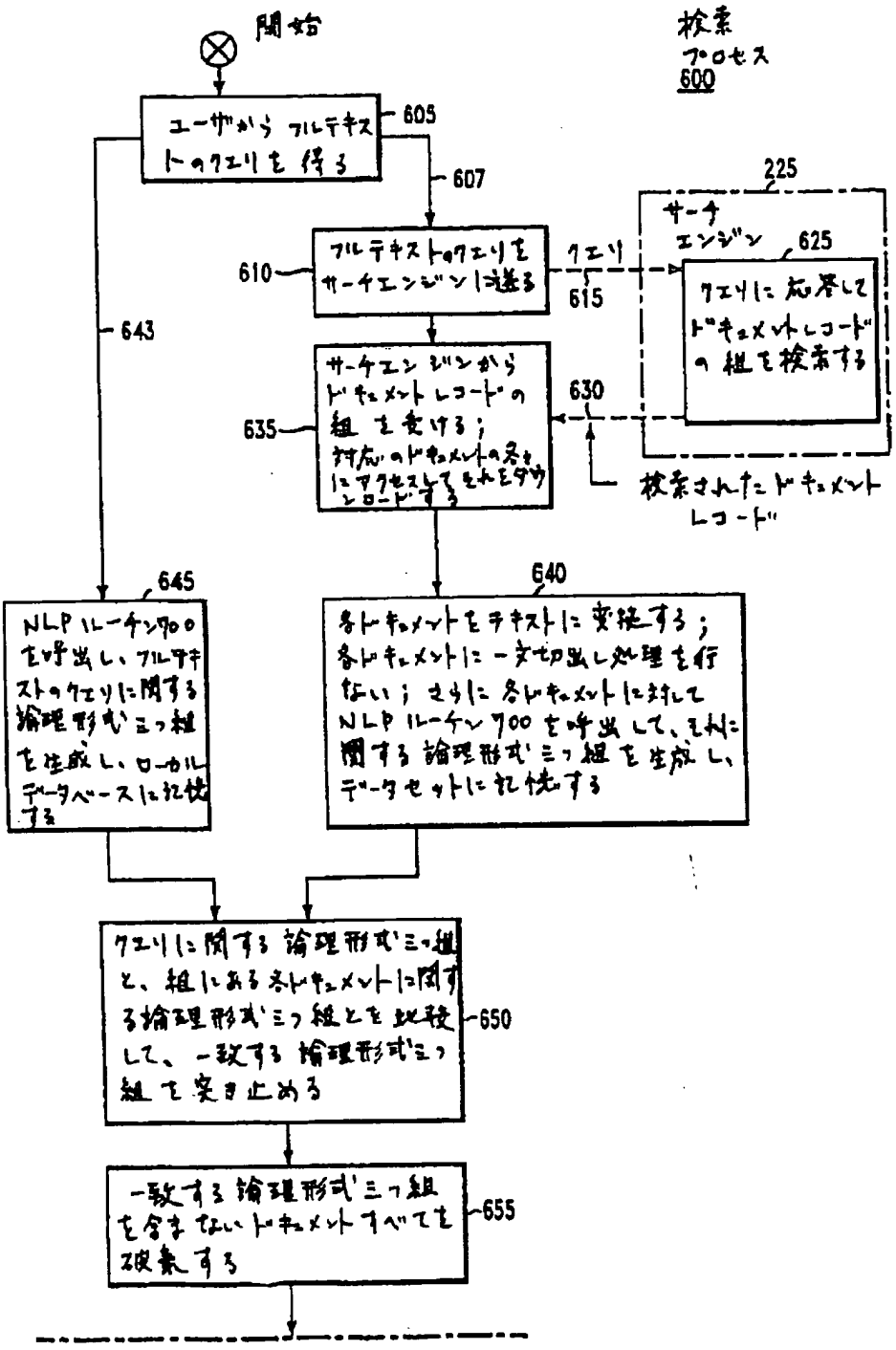
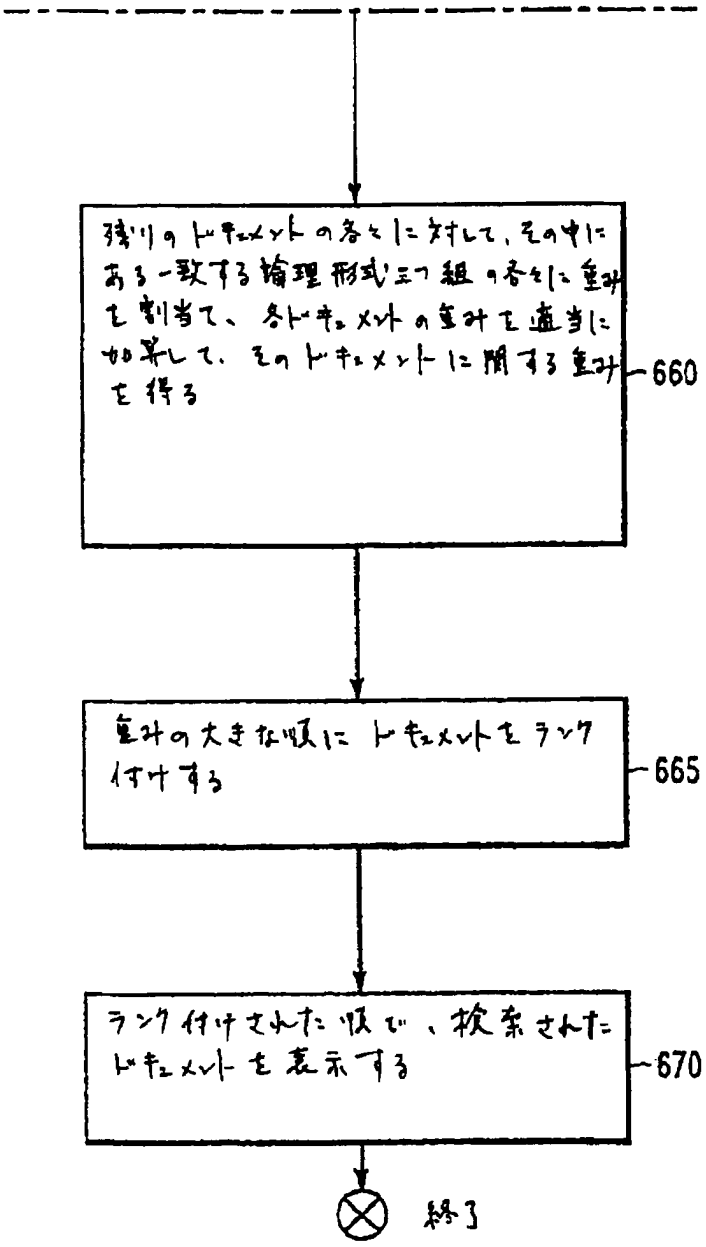


FIG. 6

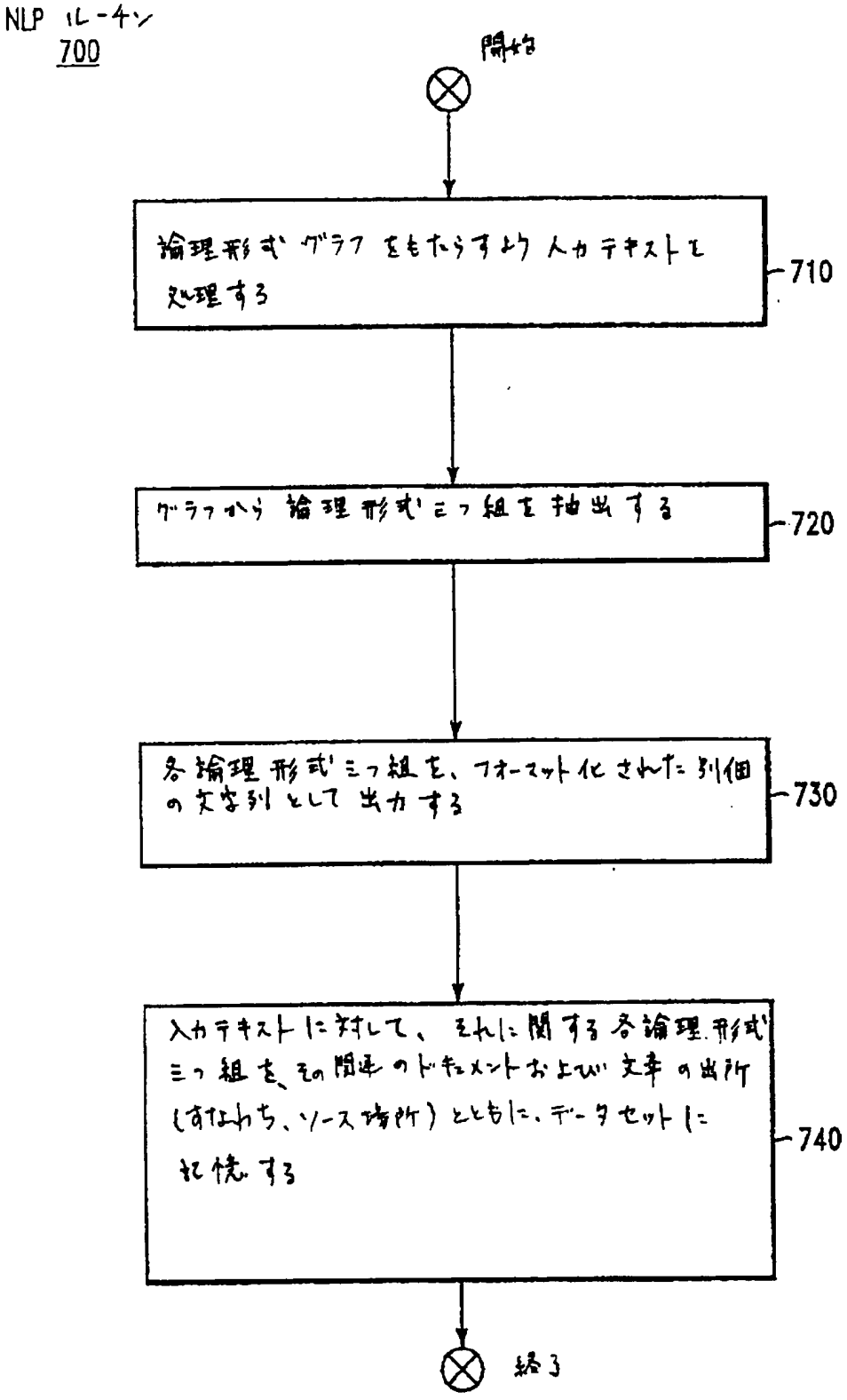
[Drawing 6 A]


開始
検索
プロセス
600
ユーザからフルテキストのクエリを得る
605
607
225
サーチエンジン
625
フルテキストのクエリをサーチエンジンに送る
610
クエリ
615
クエリに応答してドキュメントレコードの組を検索する
643
サーチエンジンからドキュメントレコードの組を受ける；対応のドキュメントの各々にアクセスしてそれをダウンロードする
630
635
検索されたドキュメントレコード
645
640
NLPルーチン700を呼出し、フルテキストのクエリに関する論理形式三つ組を生成し、ローカルデータベースに記憶する
各ドキュメントをテキストに変換する；各ドキュメントに一文切出し処理を行ない；さらに各ドキュメントに対してNLPルーチン700を呼出して、それに関する論理形式三つ組を生成し、データセットに記憶する
クエリに関する論理形式三つ組と、組にある各ドキュメントに関する論理形式三つ組とを比較して、一致する論理形式三つ組を突き止める
650
一致する論理形式三つ組を含まないドキュメントすべてを破棄する
655

[Drawing 6 B]


残りのドキュメントの各々に対して、その中にある一致する論理形式三つ組の各々に重みを割当て、各ドキュメントの重みを適当に加算して、そのドキュメントに関する重みを得る
660
重みの大きな順にドキュメントをランク付けする
665
ランク付けされた順で、検索されたドキュメントを表示する
670
終了

[Drawing 7]

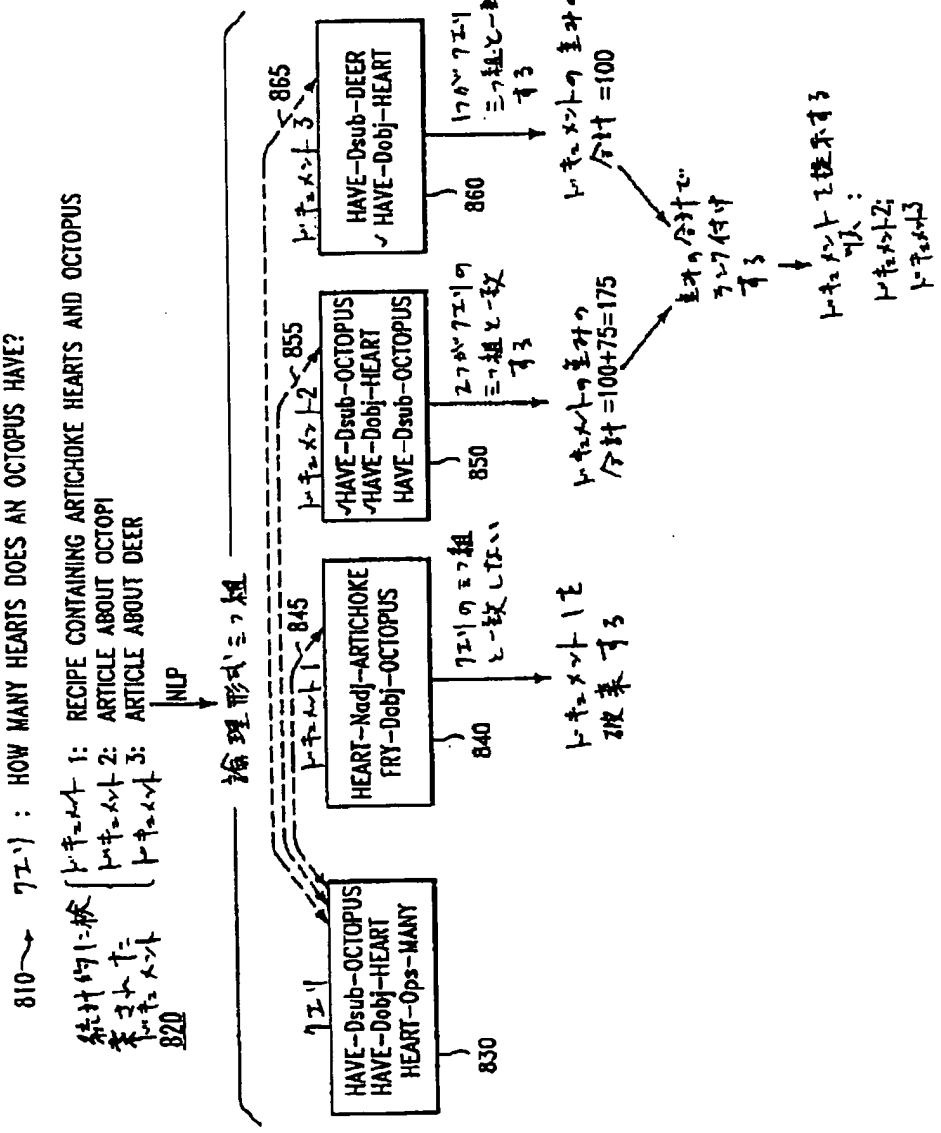

NLP ルーチン
700
開始
論理形式グラフをもたらす入力テキストを処理する
710
グラフから論理形式三つ組を抽出する
720
各論理形式三つ組を、フォーマット化された別個の文字列として出力する
730
入力テキストに対して、それに関する各論理形式三つ組を、その関連のドキュメントおよび文章の出所（すなわち、ソース場所）とともに、データセットに記憶する
740
終了

[Drawing 8 A]

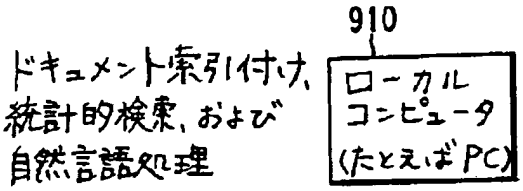
一致するミつ組の関係子のタイプ	重み
Dobj	100
Dsub	75
Ops	10
Nodj	10

800 一致論理形式ミつ組重み付け表

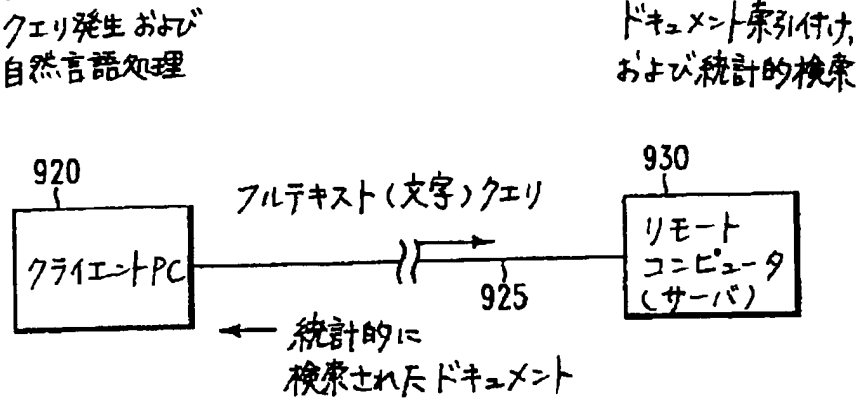
[Drawing 8 B]



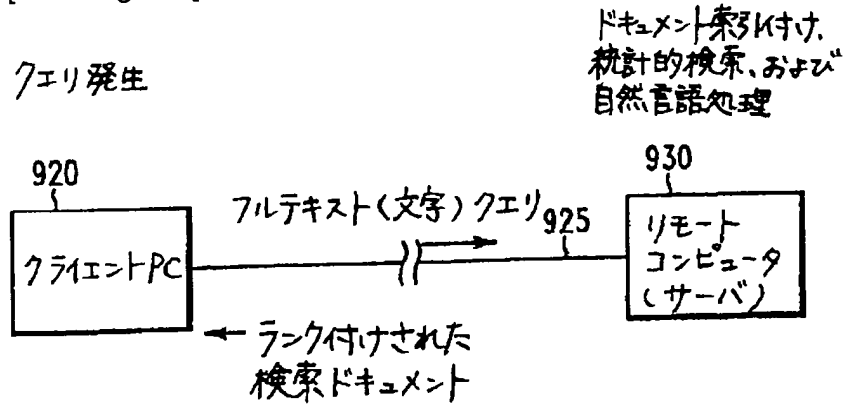
[Drawing 9 A]



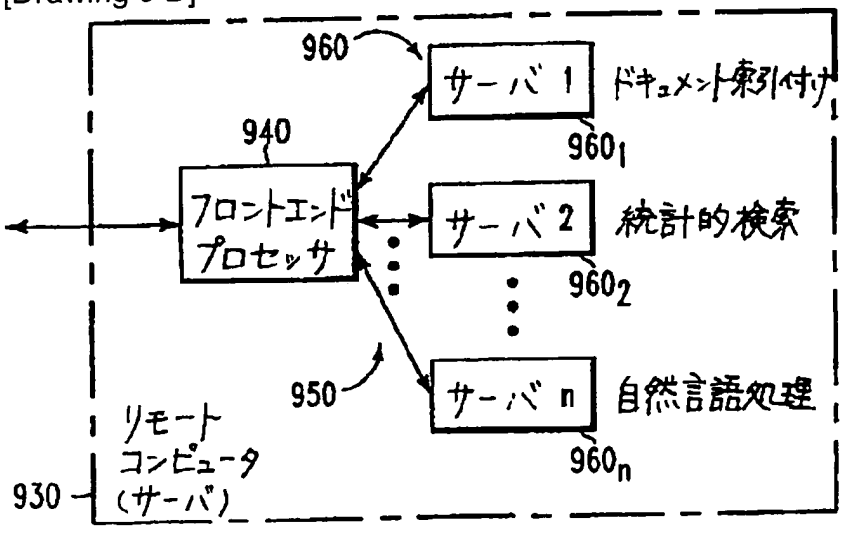
[Drawing 9 B]



[Drawing 9 C]

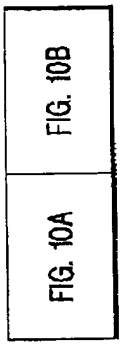


[Drawing 9 D]

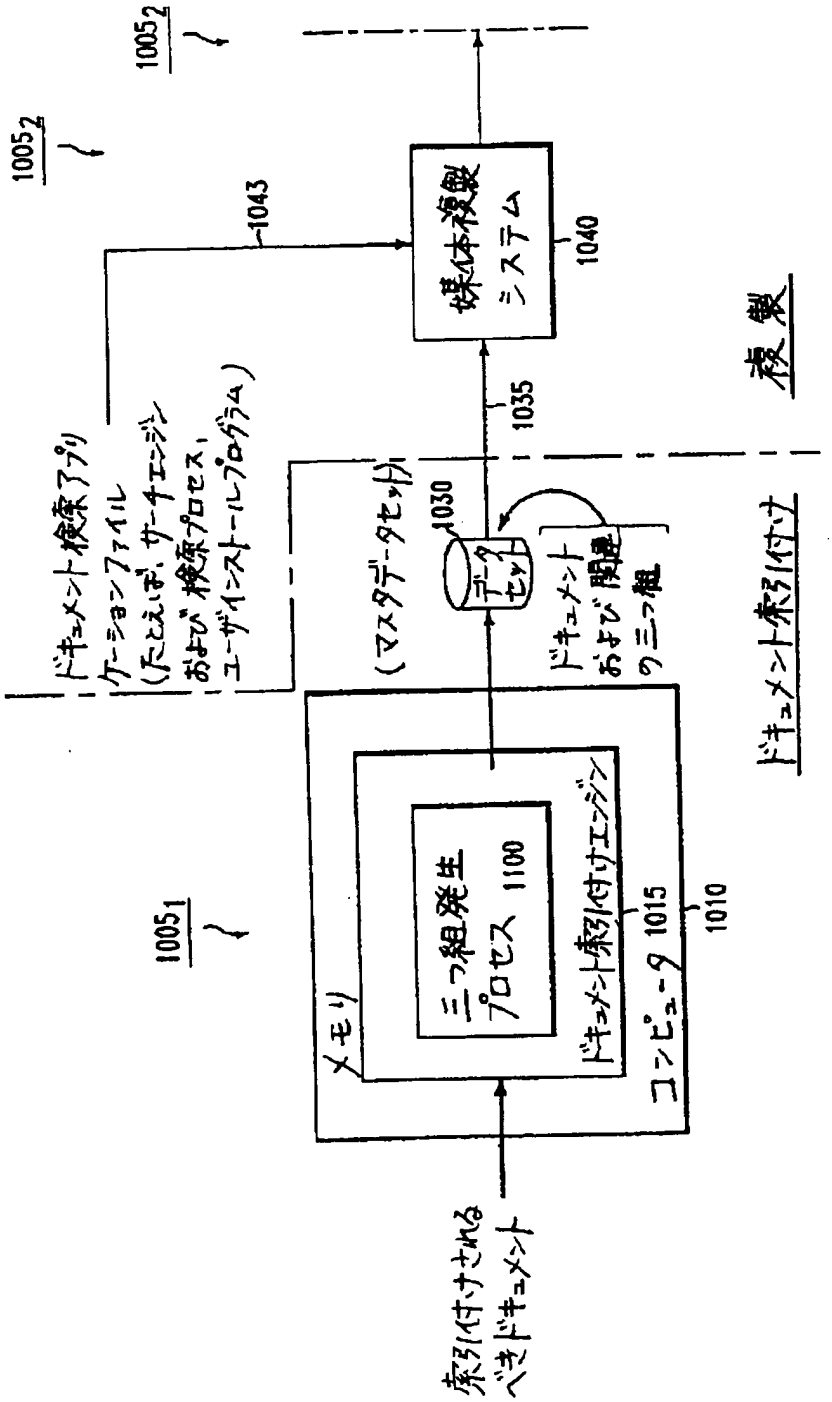


[Drawing 10]

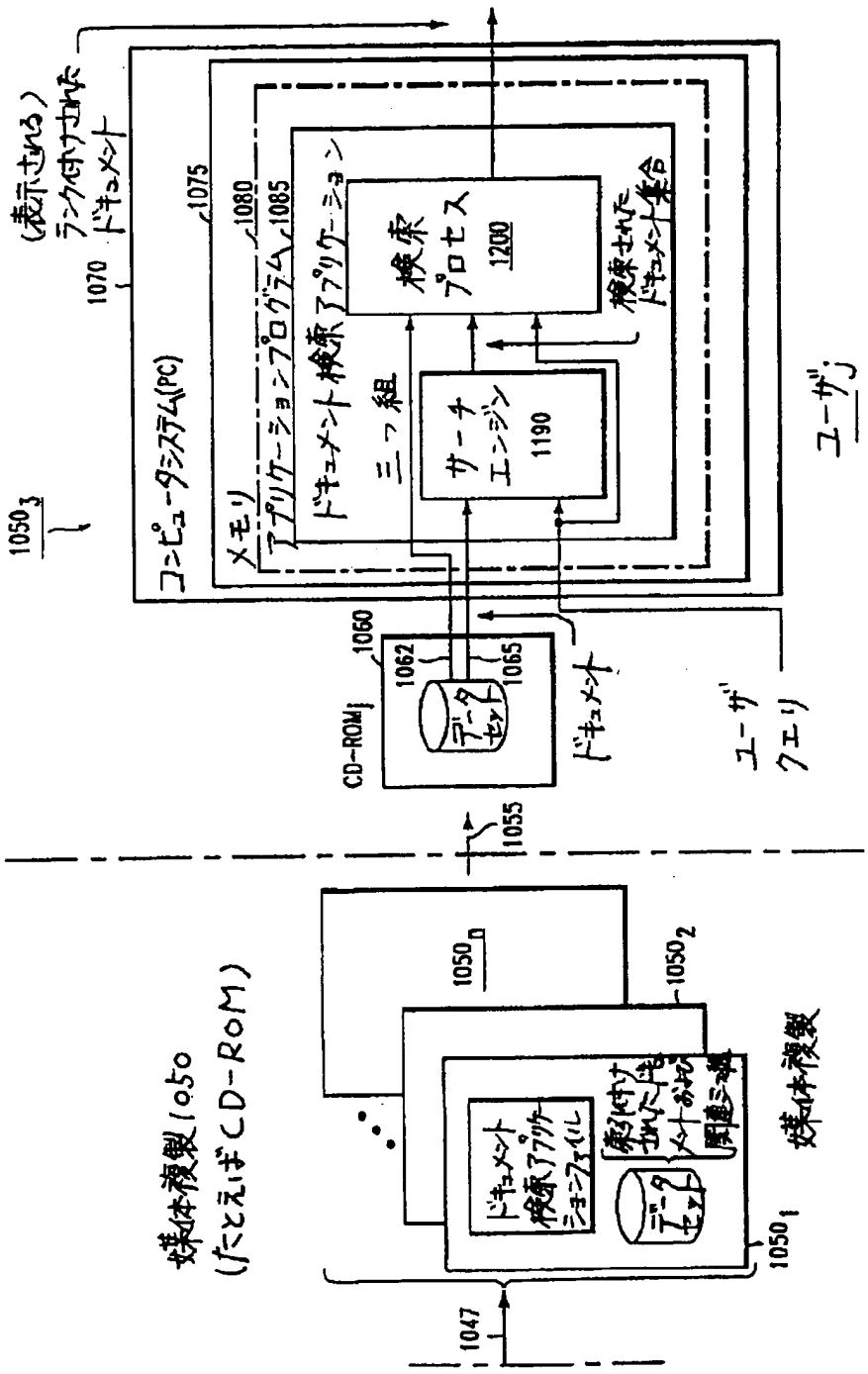
FIG. 10



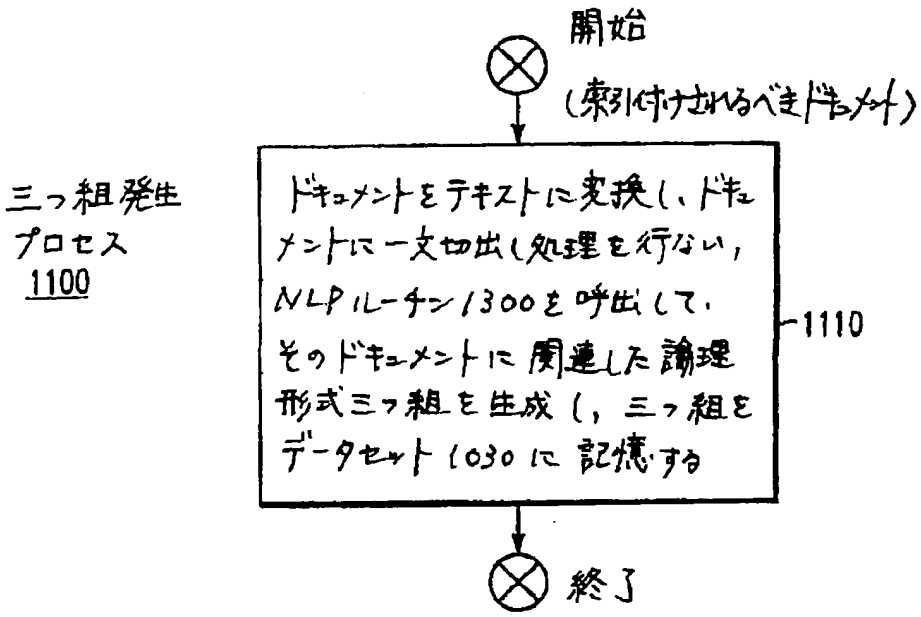
[Drawing 10 A]


1005 2
1005 2
1043
媒体複製システム
1040
ドキュメント検索アプリケーション
(たとえば、サーチエンジンおよび検索プロセス、ユーザインストールプログラム)
1035
複製
(マスタデータセット)
1030
データセット
ドキュメントおよび関連のミラー組
ドキュメント索引付け
1005 1
メモリ
三つ組発生プロセス 1100
ドキュメント索引付けエンジン
コンピュータ 1015
1010
索引付けされるべきドキュメント

[Drawing 10 B]


(表示されるランク付けされたドキュメント)
1070
1075
1080
1085
コンピュータシステム(PC)
メモリ
アプリケーションプログラム
ドキュメント検索アプリケーション
三つ組
サーチエンジン
1190
検索プロセス
1200
検索されたドキュメント集合
1050₃
ユーザ
1060
1062
1065
CD-ROM
データセット
ドキュメント
ユーザクエリ
1055
媒体複製1050
(たとえばCD-ROM)
1050n
1050₂
ドキュメント検索アプリケーションファイル
データセット
1050₁
媒体複製
1047

[Drawing 11]



[Drawing 12]

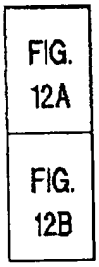
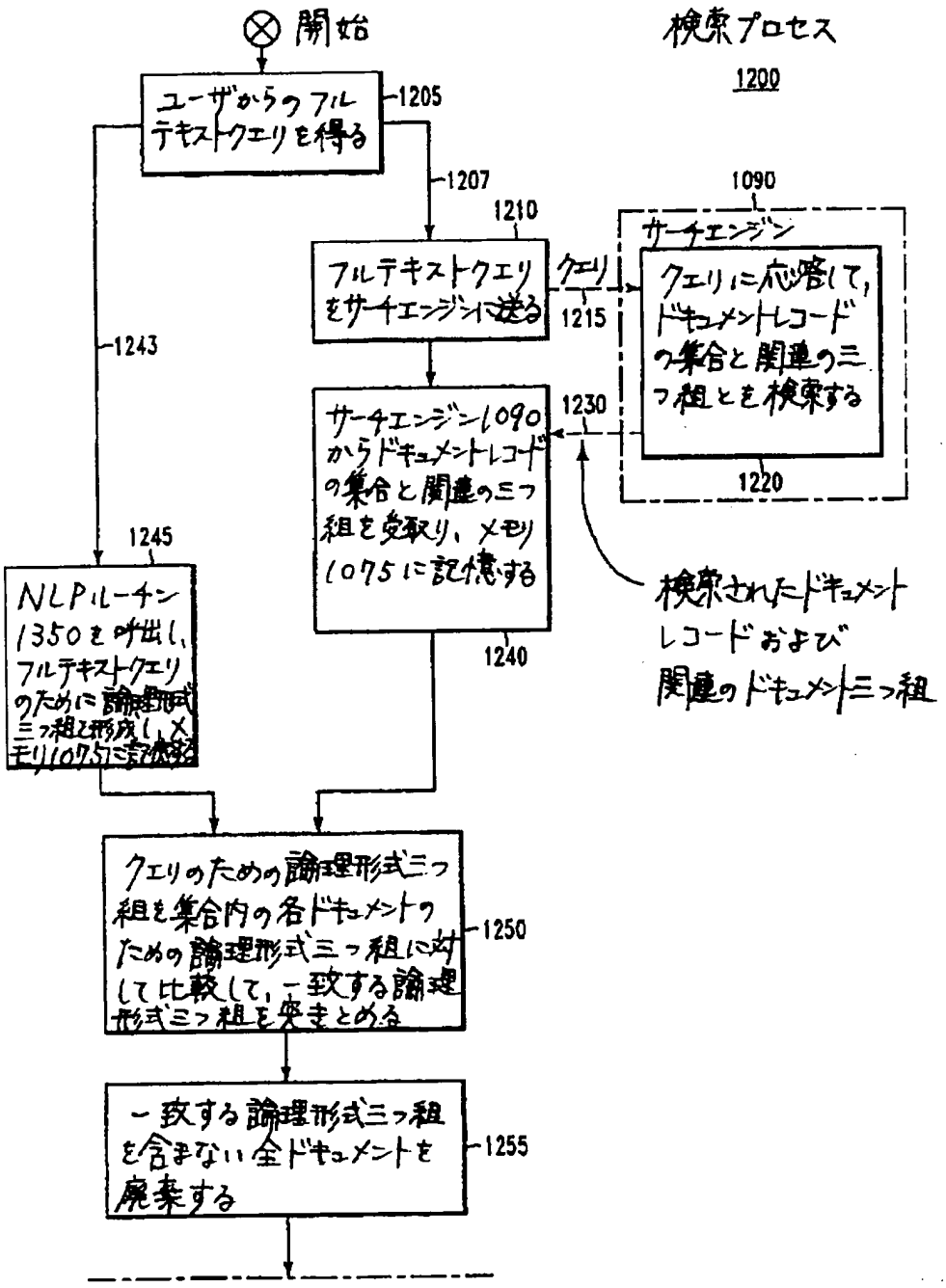
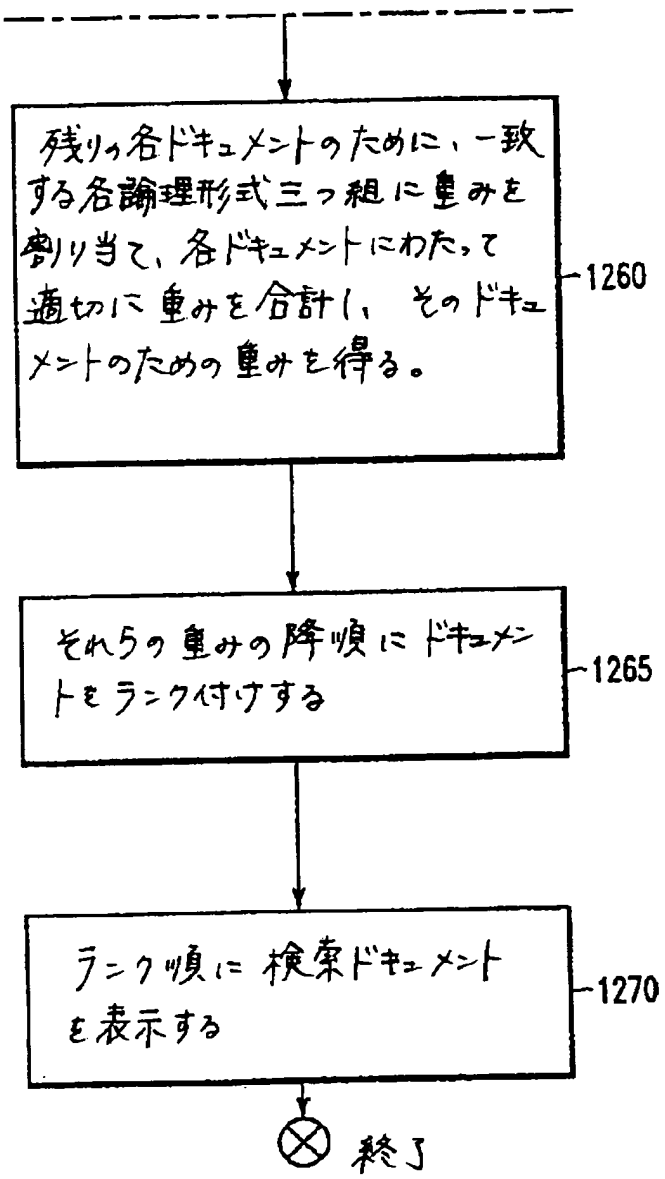


FIG. 12

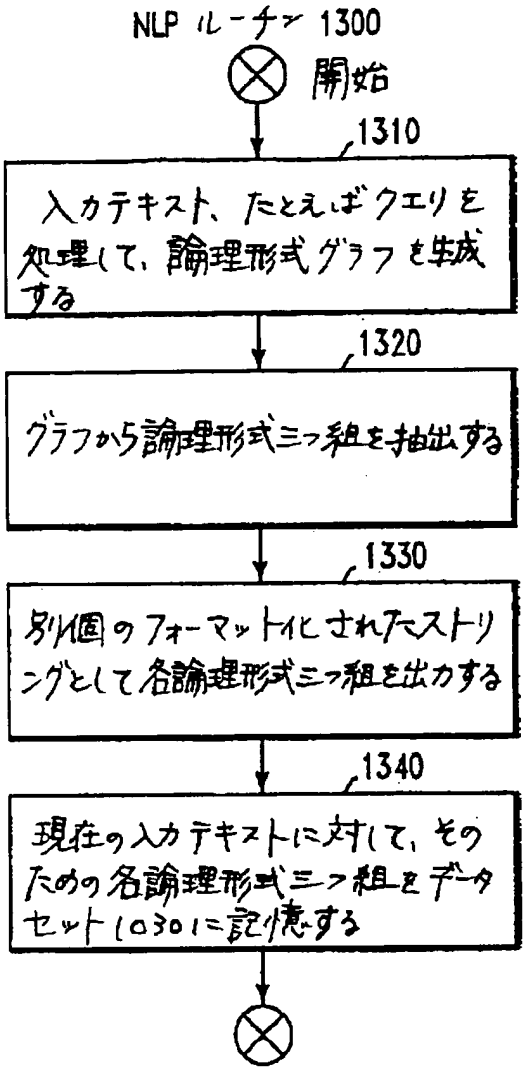
[Drawing 12 A]


開始
検索プロセス
1200
ユーザからのフルテキストクエリを得る
1205
1207
1210
フルテキストクエリをサーチエンジンに送る
クエリ
1215
1090
サーチエンジン
クエリに応答して、ドキュメントレコードの集合と関連の三つ組とを検索する
1220
1230
サーチエンジン1090からドキュメントレコードの集合と関連の三つ組を受取り、メモリ1075に記憶する
1240
検索されたドキュメントレコードおよび関連のドキュメント三つ組
1243
1245
NLPルーチン1350を呼出し、フルテキストクエリのために論理形式三つ組を形成し、メモリ1075に記憶する
クエリのための論理形式三つ組を集合内の各ドキュメントのための論理形式三つ組に対して比較して、一致する論理形式三つ組を集まとめる
1250
一致する論理形式三つ組を含まない全ドキュメントを廃棄する
1255

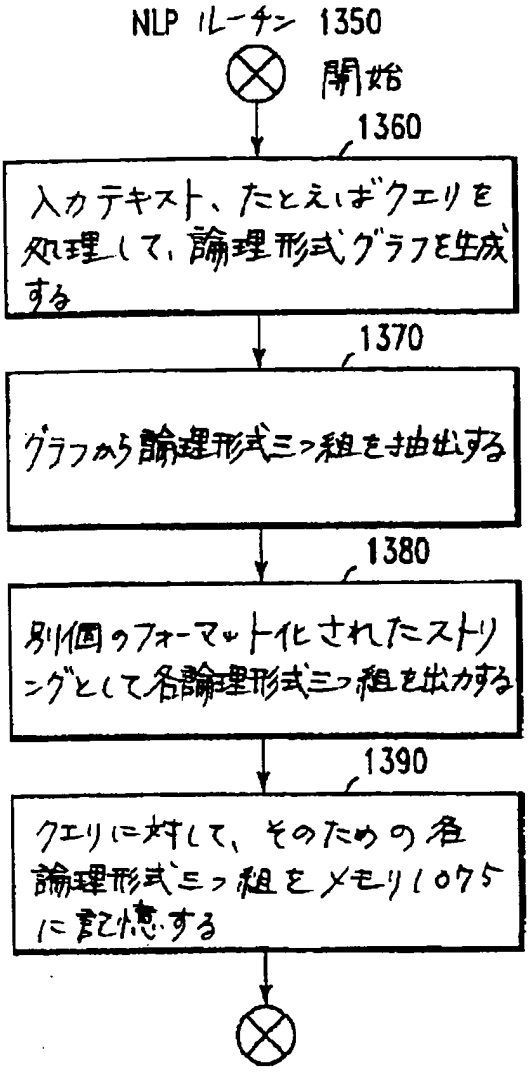
[Drawing 12 B]


残りの各ドキュメントのために、一致する各論理形式三つ組に重みを割り当て、各ドキュメントにわたって適切に重みを合計し、そのドキュメントのための重みを得る。
1260
それらの重みの降順にドキュメントをランク付けする
1265
ランク順に検索ドキュメントを表示する
1270
終了

[Drawing 13 A]


NLP ルーチン 1300
開始
1310
入力テキスト、たとえばクエリを処理して、論理形式グラフを生成する
1320
グラフから論理形式三つ組を抽出する
1330
別個のフォーマット化されたストリングとして各論理形式三つ組を出力する
1340
現在の入力テキストに対して、そのための各論理形式三つ組をデータセット1030に記憶する

[Drawing 13 B]



[Translation done.]